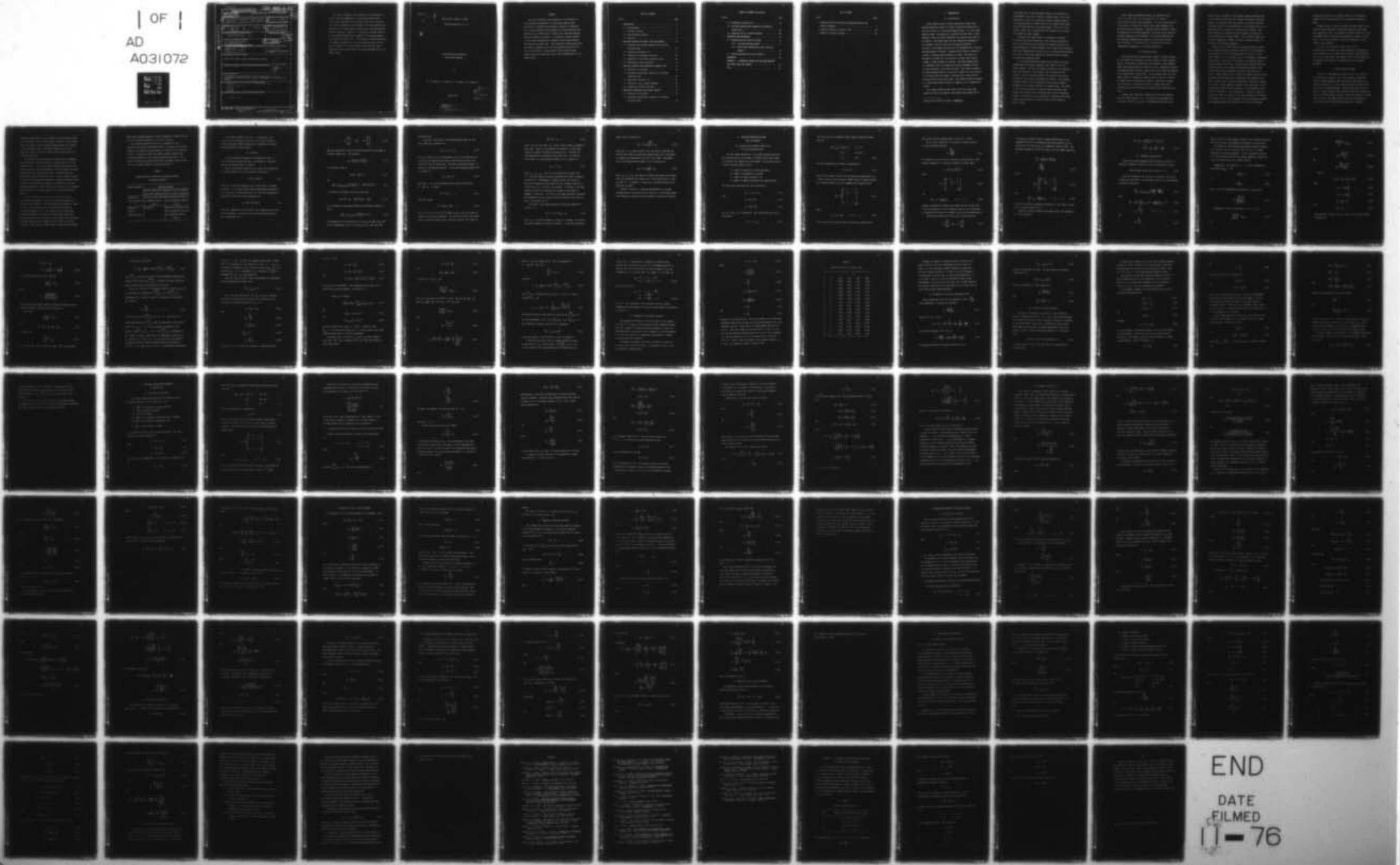
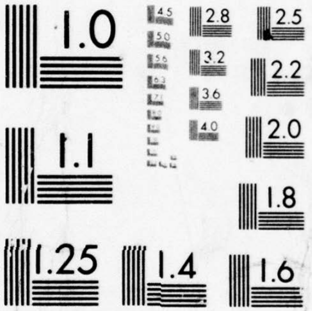
AD-A031 072    TEXAS A AND M UNIV  COLLEGE STATION    F/G 12/1
A SUPER-POPULATION APPROACH TO MULTI-STAGE SAMPLING.(U)
AUG 76  R K BURDICK, H O HARTLEY, L J RINGER    DAHC04-74-C-0018
UNCLASSIFIED    TR-17    ARO-8049.16-M    NL

| OF |
AD
A031072



END
DATE
FILMED
11-76



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ARO-8049.16-M FG.

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

12

1. ~~REPORT NUMBER~~ 19
18 ARO-8Ø49.16-M

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. ~~TITLE (and Subtitle)~~
6 A SUPER-POPULATION APPROACH TO MULTI-STAGE SAMPLING.

5. ~~TYPE OF REPORT & PERIOD COVERED~~
9 Technical Report.

6. ~~PERFORMING ORG. REPORT NUMBER~~

7. ~~AUTHOR(s)~~
10 R. K. Burdick,
H. O. Hartley
L. J. Ringer
R. L. Sielken Jr

8. ~~CONTRACT OR GRANT NUMBER(s)~~
15 DAHCØ4-74-C-0018

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Texas A&M University
College Station, Texas 77840

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
12 94p.

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. ~~REPORT DATE~~
11 August 1976

13. ~~NUMBER OF PAGES~~
87

14. ~~MONITORING AGENCY NAME & ADDRESS~~(if different from Controlling Office)
14 TR-17

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Statistics
Population (statistics)
Estimating
Sampling
Surveys
Stochastic processes

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(see other side)

347350

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

This report develops a new technique for the estimation of finite population parameters in a multi-stage sample survey. Specifically, estimators and confidence intervals for parameters of the finite population are developed for two-stage sampling when primaries are of either equal or unequal size, two-stage sampling when the variable of interest, y, is related to another variable, x, and p-stage sampling with an example of three-stage sampling when units are of equal size. The stochastic procedure generating the sample is assumed to be a two step procedure where the first step is selection of a "large sample" from an infinite superpopulation and the second step is the actual implementation of the sample survey.

ARO-D PROJECT DAHCO4 74 C 0018

Technical Report No. 17 ✓

# A SUPER-POPULATION APPROACH TO MUTLI-STAGE SAMPLING

by

R. K. Burdick, H. O. Hartley, L. J. Ringer, R. L. Sielken Jr.

August 1976

# ABSTRACT

This report develops a new technique for the estimation of finite population parameters in a multi-stage sample survey. Specifically, estimators and confidence intervals for parameters of the finite population are developed for two-stage sampling when primaries are of either equal or unequal size, two-stage sampling when the variable of interest, y, is related to another variable, x, and p-stage sampling with an example of three-stage sampling when units are of equal size. The stochastic procedure generating the sample is assumed to be a two step procedure where the first step is selection of a "large sample" from an infinite super-population and the second step is the actual implementation of the sample survey.

## TABLE OF CONTENTS

Section Page

TABLE OF CONTENTS (continued)

## LIST OF TABLES

# 1. INTRODUCTION

## 1.1 Preliminaries

In many sample surveys of finite populations, single-stage sampling designs are either too costly or physically impractical to use, and the need for a multi-stage design arises. In a multi-stage sampling design, the population is divided into large units, called primaries, which are further subdivided into smaller units known as secondaries. The primary units are then sampled and subsamples of the secondary units are taken from the selected primaries. If necessary, the secondary units may also be subsampled until a desired sampling element is obtained. Sukhatme [1947] offers an example of a survey to estimate wheat production in India. The district of Moradabad is divided into six divisions and within each division a sample of eight villages is selected. Two wheat growing fields are subsampled from each village and a further subsample of plots is chosen from each of the selected fields. Kish [1952] considers a two-stage sampling design of a city with a sample of blocks selected in the first stage and a sample of dwelling units taken from the selected blocks in the second stage. Kish [1965] provides an example of a three-stage design and gives an extensive list of similar case studies.

Both Deming [1950] and Kish [1965] state that multi-stage designs are often less expensive than single-stage designs due to

Citations will follow the format of Biometrics.

the reduced cost in preparing sample frames and the reduction of interviewers' travel time. Whereas a sampling frame of the entire population is needed for single-stage designs, sampling frames in a multi-stage design are needed only for the elements whose larger units have been selected at an earlier stage. The cost of interviewers' travel time is reduced since the elements an interviewer must sample are closer together.

The goal of any sample survey is estimation of parametric functions of the finite population. Under classical frequentist theory, any inference made concerning the population reflects the expected behavior of repeated samples from the same finite population. Under this assumption and threugh the use of conditional expectations, the theory of multi-stage sampling has developed from the results of single-stage sampling theory.

In a two-stage design when primaries are selected with equal probabilities and without replacement, Raj [1968, p. 114] and Cochran [1963, p. 304] give an unbiased estimator of the population total, the variance of the estimator, and an estimator of the variance. For the case when primaries are selected with unequal probabilities and without replacement, Raj [1968, p. 118] provides a general estimator for the population total provided unbiased estimators of the primary total and its variance exist. Raj [1968, p. 119] and Cochran [1963, p. 305] give similar estimators when primaries are selected with replacement under various subsampling schemes. Other results such as extension to stratified multi-stage sampling and estimation of ratios are found in Raj [1968] and Cochran [1963].

Unlike classical frequentist theory, the approach in this report will assume that the finite population is a sample of size $M_o$ from an infinite super-population. The actual implementation of the survey is a process of taking a multi-stage subsample from the selected finite population. Inferences may be desired for either the parameters of the super-population, or for the finite population, although technically the "parameters" of the finite population are now statistics summarizing the sample of size $M_o$ drawn from the super-population. This report will be concerned with inferences regarding the "parameters" of the finite population.

## 1.2 Literature Review

Before discussing the super-population model, it seems necessary to first mention the direction of recent research in survey sampling theory. Rao [1971] has stated that until recently, survey sampling theory has evolved through an inductive process. Reasonable estimators and sampling designs have been suggested and their properties examined either analytically or empirically. It has become evident that some general theory of sampling is needed to better relate the inference of sample surveys to statistical inference. Many recent attempts have been made to formalize the theory of sampling from finite populations, but much confusion and controversy are associated with these.

Godambe [1955, 1966] has introduced the notion of estimators which are label dependent, i.e., not invariant to permutations of the labels attached to the units. He has shown that in this more general class of estimators, the customary optimal estimators used in the classical theory lose their classical optimality properties. However, the estimators which are superior to the label invariant estimators depend on the characteristics of non-sampled units in the population and are, therefore, not practically available. Many other authors have extended and refined Godambe's original work and are referenced in Godambe [1969]. Ericson [1969a, 1969b] used the model suggested by Godambe to consider a Bayesian approach of inference for the finite population.

Hartley and Rao [1968, 1969] and independently Royall [1968] have established certain optimality properties for the customarily used estimators in the basic designs within the class of "scale load" estimators, i.e., estimators which are invariant to permutations of the labels. Practically all estimators that have been used in practice belong to this class, although Hartley and Rao [1969] have stated that they do not exclude label dependent estimators from consideration and give examples of instances when they will be useful. However, they have not provided any general guidelines which infallibly indicate under what circumstances label dependent estimators should be used. They have stated that a sufficient condition for the use of label dependent estimators arises when some or all of the parameter functions in the population to be estimated are themselves not invariant to label permutations. Among the results of this theory within the class of "scale load" estimators are UMV-ness of the sample mean in random sampling, the Horvitz-Thompson estimator when sampling probability proportional to size, and the maximum likelihood properties of an estimator similar to the regression estimator when the population mean of a concomitant variable is known.

Godambe and Sprott [1971] and Johnson and Smith [1969] contain excellent papers expressing the different viewpoints concerning these sampling theories. The comments of Barnard [1969], Rao [1971], and Godambe [1970] also indicate the diversity of opinion on these issues. Rao [1973] provides an extensive bibliography of other studies on this problem.

In using a super-population model, this report is concerned with a new theory for survey sampling. By assuming the finite population to be a sample from an infinite super-population defined by a linear model, the general theory of linear models can be applied to the problem of estimating the finite population parameters.

## 1.3 Super-Population Models

The use of a super-population model is not a new concept in the statistical literature. One of the earliest to explicitly state the super-population model was Cochran [1939] when he considered estimation of the finite population mean for simple random and stratified sampling designs. Cochran [1946] again used the model to compare systematic and stratified samples from populations where the variance within a group of elements increases as the group size increases. Raj [1958] regarded the finite population as a random sample from an infinite super-population to compare a probability proportional to size estimator with the simple average, ratio, regression, and stratified sample estimators. Royall [1970] used such a model to develop optimal sampling plans for estimating a finite population total. Many other examples exist and the interested reader is referred to Fuller [1973] and Rao [1973] for additional references.

The notion of an underlying super-population also accompanied the introduction of analytical surveys. Deming [1950, Ch. 7] and Cochran [1963, p. 37] state that when comparing domain means in an analytical survey, the null hypothesis is that the domains have been drawn from the same infinite population. Sedransk [1965] also expresses the view that inferences refer to a more "general" population than the existing finite population. Konijn [1962] made similar assumptions and considered estimators for functions of the super-population parameters. Fuller [1973] gives results for the estimation of parameters of the infinite population in a two-stage sampling design.

Of more particular interest in this report are the papers concerned with estimation of the finite population parameters as opposed to the estimation of the super-population parameters. Royall and Herson [1973a, 1973b] assumed a super-population model in estimating parameters of the finite population for single-stage designs and examined results when the assumed model broke down. Hartley and Sielken [1975] considered a more general case than Royall and Herson where auxiliary variables are not fixed in the super-population. Scott and Smith [1969] assumed a super-population model when using a Bayesian approach to derive estimators of linear functions of finite population parameters in two-stage sampling.

The preceding discussion may best be summarized by Table 1 reproduced from Hartley and Sielken [1975]. In regard to multi-stage sampling, the standard results discussed in Section 1.1 are classified as Case 1. Analytical surveys and similar studies concerned with estimating parameters of the infinite population are Case 3. The papers of Hartley and Sielken [1975], Royall and Herson [1973a, 1973b], and Scott and Smith [1969] are concerned with Case 2.

TABLE 1

Sampling Theories Classified by Sampling Procedure and Target Parameters

| Target Parameters | Sampling Procedure | |
|---|---|---|
| | Repeated sampling from a fixed finite population | Repeated two-step sampling from an infinite population |
| Parameters of finite population | Classical finite population sampling theory<br>= Case 1 | Super-population theory for finite population sampling<br>= Case 2 |
| Parameters of infinite super-population | Infeasible | Inference on infinite population parameters from two-step sampling procedure<br>= Case 3 |

This report considers only Case 2. In particular, this report takes a non-Bayesian approach to the problem considered by Scott and Smith [1969] and serves as an extension to the work done by Hartley and Sielken [1975].

### 1.4 Overview

In this section the problems to be considered in this report are briefly sketched, and their relationship to the work of Hartley and Sielken [1975] is discussed.

Hartley and Sielken assume that the current finite population is a random sample from a super-population of the form

$$y = \underline{x}^T\underline{\beta} + \varepsilon[v(x)]^{\frac{1}{2}} \qquad (1.1)$$

where $\underline{x}$ is a vector of variables, $\underline{\beta}$ is a (p×1) vector of unknown constants, $\varepsilon$ is a normal random variable independently distributed of x, and v(x) is any known function of x. The basic parameters of interest for the current finite population are

$$\underline{b} = (X^TV^{-1}X)^{-1}X^TV^{-1}\underline{Y} \qquad (1.2)$$

and linear combinations of $\underline{b}$, say $\underline{c}^T\underline{b}$. Once a sample survey of the finite population is conducted, the population quantities $\underline{Y}$ and X can be partitioned into

$$\underline{Y} = \begin{bmatrix} \underline{Y}_s \\ \\ \underline{Y}_r \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} X_s \\ \\ X_r \end{bmatrix} \tag{1.3}$$

where the subscripts "s" and "r" indicate inclusion in the sample or exclusion respectively. The quantity

$$\underline{b}_s = (X_s^T V_s^{-1} X_s)^{-1} X_s^T V_s^{-1} \underline{Y}_s \tag{1.4}$$

is such that for any $\underline{c}$,

$$E(\underline{c}^T \underline{b} - \underline{c}^T \underline{b}_s) = 0 \tag{1.5}$$

and

$$\underline{c}^T \underline{b}_s \pm \hat{\sigma}_s t_{n-p;\alpha/2} \{\underline{c}^T [(X_s^T V_s^{-1} X_s)^{-1} - (X^T V^{-1} X)^{-1}] \underline{c}\}^{\frac{1}{2}} \tag{1.6}$$

is a 100(1-α)% "confidence interval" on $\underline{c}^T \underline{b}$ where

$$(n-p)\, \hat{\sigma}_s^2 = (\underline{Y}_s - X_s \underline{b}_s)^T V_s^{-1} (\underline{Y}_s - X_s \underline{b}_s) \ . \tag{1.7}$$

If X is unknown, an approximate 100(1-α)% "confidence interval" on $\underline{c}^T \underline{b}$ is

$$\underline{c}^T \underline{b}_s \pm \hat{\sigma}_s t_{n-p;\alpha/2} \{\underline{c}^T (X_s^T V_s^{-1} X_s)^{-1} \underline{c}\}^{\frac{1}{2}} \ . \tag{1.8}$$

No assumptions on the distributions of X and $X_s$ are made except that they are independent of the ε's and that $X_s$ is of full rank with probability one.

In Section 2 and Section 3 the super-population model for two-stage sampling is assumed to be

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{1.9}$$

where $y_{ij}$ refers to the j-th observation in the i-th primary and the $\alpha_i$ and $\varepsilon_{ij}$ are independently normally distributed with means 0 and variances $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$ respectively. The finite population parameter of interest is the population mean $\bar{Y}$. If the super-population model (1.9) is rewritten as

$$y_{ij} = \alpha_i^* + \varepsilon_{ij} \tag{1.10}$$

where $\alpha_i^* = \mu + \alpha_i$, then conditioning upon the $\alpha_i^*$'s and letting $\underline{\beta}^T = [\alpha_1^* \ \ \alpha_2^* \ \ \ldots]$ gives

$$\underline{b}^T = [\bar{Y}_1 \ \bar{Y}_2 \ \ \ldots] \ . \tag{1.11}$$

Then $\underline{c}^T\underline{b} = \bar{Y}$ when

$$\underline{c}^T = [M_1/M_o \ \ M_2/M_o \ \ \ldots] \tag{1.12}$$

where $M_i$ is the size of the $i^{th}$ primary and $M_o$ is the total number of elements in the finite population. The results of Hartley and Sielken imply an unbiased estimator and a confidence interval for $\bar{Y}$ based on

$$\underline{b}_s^T = [\bar{y}_1 \;\; \bar{y}_2 \;\; \ldots] \tag{1.13}$$

only if $X_s$ is of full rank, i.e., only if every primary is sampled at least once. Since not all primaries are sampled in a multi-stage design, the results of Hartley and Sielken do not apply when the super-population is considered in the form (1.10). To alleviate this problem, the super-population model (1.9) is rewritten as

$$y_{ij} = \mu + \eta_{ij} \tag{1.14}$$

where $\eta_{ij} = \alpha_i + \varepsilon_{ij}$. Then, if the primaries are of equal size, $b = \bar{Y}$, and the results of Hartley and Sielken will imply an unbiased estimator of $\bar{Y}$. Furthermore, if $\sigma_\alpha^2/\sigma_\varepsilon^2$ is known, the results of Hartley and Sielken will also imply an exact confidence interval on $\bar{Y}$ since then X, $X_s$, V, and $V_s$ are all known. In Section 2, two-stage sampling in which the primaries are of equal size but $\sigma_\alpha^2/\sigma_\varepsilon^2$ is <u>unknown</u> is considered. In Section 3, two-stage sampling is considered when primaries are not of equal size and consequently there does not exist a $\underline{c}$ such that $\underline{c}^T\underline{b} = \bar{Y}$.

In Section 4, the super-population for two-stage sampling is assumed to be

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij} \tag{1.15}$$

where $x_{ij}$ is a variable related to y and β is a constant. The finite population parameter of interest is still $\bar{Y}$. If the super-population model (1.15) is rewritten as

$$y_{ij} = [1 \ x_{ij}] \begin{bmatrix} \alpha_i^* \\ \beta \end{bmatrix} + \varepsilon_{ij} \qquad (1.16)$$

where $\alpha_i^* = \mu + \alpha_i$, then as with (1.10), the results of Hartley and Sielken do not apply to multi-stage sampling since not all primaries are sampled and consequently $X_s$ is not of full rank. Furthermore, even if the super-population model (1.15) is rewritten as

$$y_{ij} = [1 \ x_{ij}] \begin{bmatrix} \mu \\ \beta \end{bmatrix} + \eta_{ij} \qquad (1.17)$$

where $\eta_{ij} = \alpha_i + \varepsilon_{ij}$, the results of Hartley and Sielken do not apply unless all primaries are of equal size so that there exists a $\underline{c}$ such that $\underline{c}^T\underline{b} = \bar{Y}$. In addition, to construct a confidence interval on $\bar{Y}$, $\sigma_\alpha^2/\sigma_\varepsilon^2$ must be known.

Finally, in Section 5, a general methodology for a p-stage sampling design is discussed and the results for a three-stage design with primaries of equal size and secondaries of equal size are given.

# 2. TWO-STAGE SAMPLING WITH EQUAL SIZES AND SAMPLES

## 2.1 Estimation and Variance Formulas for the Finite Population Mean

The first model considered is a two-stage sampling design when all primaries have an equal number of elements and an equal number of secondaries are sampled from each primary. The notation adopted is that of Cochran [1963] in which

$N$ = number of primaries in finite population,

$M$ = number of secondaries per primary,

$n$ = number of sampled primaries, and

$m$ = number of sampled secondaries per sampled primary.

The linear model describing the super-population is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{2.1}$$

where

$$\alpha_i \sim N(0, \sigma_\alpha^2) \ , \tag{2.2}$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \ , \tag{2.3}$$

and all $\alpha_i$ and $\varepsilon_{ij}$ are independent. This linear model may also be expressed as

$$y_{ij} = \mu + \eta_{ij} \tag{2.4}$$

where the $\eta_{ij}$'s are independent normal random variables with mean zero and

$$\begin{aligned} E(\eta_{ij}\ \eta_{k\ell}) &= \sigma_\alpha^2 + \sigma_\varepsilon^2 , && i=k,\ j=\ell , \\ &= \sigma_\alpha^2 , && i=k,\ j\neq\ell , \\ &= 0 , && i\neq k . \end{aligned} \tag{2.5}$$

The finite population of size MN is represented by

$$\underline{Y} = \underline{1}\mu + \underline{H} \tag{2.6}$$

where $\underline{Y}$ is the (MN×1) vector of finite population observations, $\underline{1}$ is a (MN×1) vector of ones, and $\underline{H}$ is a (MN×1) vector of random variables. The covariance matrix of $\underline{H}$ is the (MN×MN) block diagonal matrix

$$\sigma_\varepsilon^2 V = \sigma_\varepsilon^2 \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & V_N \end{bmatrix} \tag{2.7}$$

where

$$V_i = I_M + \rho J_M , \qquad i = 1, \ldots, N , \tag{2.8}$$

with $I_M$ denoting the identity matrix of order M, $J_M$ denoting the

(M×M) matrix with all elements equal to one, and $\rho = \sigma_\alpha^2/\sigma_\varepsilon^2$ .

For the finite population, the quantity of present interest is the finite population mean

$$\bar{Y} = \frac{\sum\limits_{ij}^{NM}\sum y_{ij}}{MN} . \qquad (2.9)$$

Alternatively $\bar{Y}$ can be written in the form of the BLUE (best linear unbiased estimator) of μ based on a sample of size MN, namely

$$\bar{Y} = (\underline{1}^T V^{-1}\underline{1})^{-1}\underline{1}^T V^{-1}\underline{Y} \qquad (2.10)$$

where

$$V^{-1} = \begin{bmatrix} V_1^{-1} & 0 & \dots & 0 \\ 0 & V_2^{-1} & \dots & 0 \\ \vdots & & & \\ 0 & . & \dots & V_N^{-1} \end{bmatrix} \qquad (2.11)$$

and

$$V_i^{-1} = I_M - \frac{\rho}{(1+M\rho)} J_M , \qquad i = 1, \dots, N . \qquad (2.12)$$

However, as discussed by Hartley and Sielken [1975], the fact that $\bar{Y}$ can be expressed as (2.10) is important only in that it motivates an estimator for $\bar{Y}$ based on the sample, and not for its BLUE property.

The population quantities $\underline{Y}$ and $\underline{1}$ can be partitioned into

$$\underline{Y} = \begin{bmatrix} \underline{Y}_s \\ \\ \underline{Y}_r \end{bmatrix} \quad \text{and} \quad \underline{1} = \begin{bmatrix} \underline{1}_s \\ \\ \underline{1}_r \end{bmatrix} \qquad (2.13)$$

in which $\underline{Y}_s$ is a (mn×1) vector of sampled observations, $\underline{Y}_r$ is a ((MN-mn)×1) vector of the unobserved population elements, $\underline{1}_s$ is a (mn×1) vector of ones, and $\underline{1}_r$ is a ((MN-mn)×1) vector of ones. The estimator analogous to (2.10) but based on only the sample quantities is

$$\hat{\bar{Y}}_E = (\underline{1}_s^T V_s^{-1} \underline{1}_s)^{-1} \underline{1}_s^T V_s^{-1} \underline{Y}_s$$

$$= \frac{\sum\limits_{ij}^{nm}\sum y_{ij}}{mn} \tag{2.14}$$

where

$$V_s^{-1} = \begin{bmatrix} V_{s1}^{-1} & 0 & \cdots & 0 \\ 0 & V_{s2}^{-1} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & V_{sn}^{-1} \end{bmatrix} \tag{2.15}$$

and

$$V_{si}^{-1} = I_m - \frac{\rho}{(1+m\rho)} J_m \ , \qquad i = 1, \ldots, n \ . \tag{2.16}$$

Since both $\bar{Y}$ and $\hat{\bar{Y}}_E$ are unbiased estimators of $\mu$, $E(\bar{Y} - \hat{\bar{Y}}_E) = 0$, and $\hat{\bar{Y}}_E$ is a natural estimator for $\bar{Y}$.

Using the results of Hartley and Sielken [1975], the variance of $(\bar{Y} - \hat{\bar{Y}}_E)$ is given by

$$V(\bar{Y} - \hat{\bar{Y}}_E) = (\underline{1}_s^T V_s^{-1} \underline{1}_s)^{-1} - (\underline{1}^T V^{-1} \underline{1})^{-1}$$

$$= \frac{\sigma_\alpha^2}{n} (1 - \frac{n}{N}) + \frac{\sigma_\varepsilon^2}{mn} (1 - \frac{mn}{MN}) \quad . \tag{2.17}$$

## 2.2 Confidence Intervals on $\bar{Y}$

Since for the super-population both $\bar{Y}$ and $\hat{\bar{Y}}_E$ are statistics, a 100(1-α)% "confidence interval" on $\bar{Y}$ will be interpreted to mean

$$\text{Prob } (\bar{Y} \varepsilon [\text{Lower bound, Upper bound}]) = 1 - \alpha \ . \tag{2.18}$$

Under the assumption that the ratio of variances ρ and hence V is known, an exact 100(1 - α)% confidence interval on $\bar{Y}$ corresponding to Hartley and Sielken is

$$\left[\hat{\bar{Y}}_E \pm \hat{\sigma}_s t_{\alpha/2;mn-1} \sqrt{\frac{1+m\rho}{mn} - \frac{1+M\rho}{MN}}\right] \tag{2.19}$$

where

$$(mn - 1)\hat{\sigma}_s^2 = \sum_{ij}^{nm}\sum (y_{ij} - \bar{y})^2 - (\frac{m\rho}{1+m\rho}) \sum_i^n m(\bar{y}_i - \bar{y})^2 \ , \tag{2.20}$$

$$\bar{y}_i = \frac{\sum_j^m y_{ij}}{m} \ , \qquad i = 1, \ldots, n \ , \tag{2.21}$$

and

$$\bar{y} = \frac{\sum_i^n \bar{y}_i}{n} \ . \tag{2.22}$$

However, since $\rho$ is rarely known in practice, two confidence intervals which do not require its knowledge are developed in this section.

One method for constructing an approximate confidence interval on $\bar{Y}$ is based on the fact that $V(\bar{Y} - \hat{\bar{Y}}_E)$ in (2.17) is a linear combination of variance components. To set a confidence interval on $\bar{Y}$, some well known results from variance components analysis are stated as theorems (see, e.g., Graybill [1961]).

Theorem 2.1: Let $n_i x_i/\sigma_i^2$ $(i = 1, \ldots, p)$ be independently distributed as $\chi^2_{(n_i)}$. Let

$$\gamma = \sum_i^p g_i \sigma_i^2 > 0 \qquad (2.23)$$

and

$$g = \sum_i^p g_i x_i \ . \qquad (2.24)$$

Then $U = n'g/\gamma$ is approximately distributed as a $\chi^2_{(n')}$ where

$$n' = \frac{(\sum_i^p g_i \sigma_i^2)^2}{\sum_i^p (g_i^2 \sigma_i^4/n_i)} \ . \qquad (2.25)$$

Theorem 2.2: Under the assumptions in (2.1) - (2.3),

$$\frac{(n-1)s_b^2}{\sigma_\varepsilon^2 + m\sigma_\alpha^2} \sim \chi^2_{(n-1)} \qquad (2.26)$$

and

$$\frac{n(m-1)s_w^2}{\sigma_\varepsilon^2} \sim \chi^2_{n(m-1)} \tag{2.27}$$

where

$$s_b^2 = \frac{\sum_i^n m(\bar{y}_i-\bar{y})^2}{n-1} \tag{2.28}$$

and

$$s_w^2 = \frac{\sum_{ij}^{nm}\Sigma(y_{ij}-\bar{y}_i)^2}{n(m-1)} \ . \tag{2.29}$$

Now let

$$\gamma = V(\bar{Y} - \hat{\bar{Y}}_E) = g_1\sigma_1^2 + g_2\sigma_2^2 \tag{2.30}$$

where

$$g_1 = (1 - \frac{n}{N}) \frac{1}{mn} \ , \tag{2.31}$$

$$g_2 = (1 - \frac{m}{M}) \frac{1}{Nm} \ , \tag{2.32}$$

$$\sigma_1^2 = m\sigma_\alpha^2 + \sigma_\varepsilon^2 \ , \tag{2.33}$$

and

$$\sigma_2^2 = \sigma_\varepsilon^2 \ . \tag{2.34}$$

Using Theorem 2.1 with $x_1 = s_b^2$, $x_2 = s_w^2$, $n_1 = n-1$, and $n_2 = n(m-1)$, it follows that

$$g = g_1 x_1 + g_2 x_2$$

$$= (1 - \frac{n}{N}) \frac{s_b^2}{mn} + (1 - \frac{m}{M}) \frac{s_w^2}{Nm} \qquad (2.35)$$

is an unbiased estimator for $V(\bar{Y} - \hat{\bar{Y}}_E)$ and

$$\frac{n'g}{V(\bar{Y}-\hat{\bar{Y}}_E)} \dot{\sim} \chi^2_{(n')} \qquad (2.36)$$

where

$$n' = \frac{(g_1\sigma_1^2+g_2\sigma_2^2)^2}{g_1^2\sigma_1^4/n_1+g_2^2\sigma_2^4/n_2} . \qquad (2.37)$$

Since $\sigma_1^2$ and $\sigma_2^2$ are usually unknown, they may be replaced by their unbiased estimators, $s_b^2$ and $s_w^2$, when solving for $n'$.

Now, since

$$\frac{n'g}{V(\bar{Y}-\hat{\bar{Y}}_E)} \dot{\sim} \chi^2_{(n')} \qquad (2.38)$$

and

$$(\bar{Y} - \hat{\bar{Y}}_E) \sim N(0, V(\bar{Y} - \hat{\bar{Y}}_E)) , \qquad (2.39)$$

it follows that

$$\frac{\bar{Y}-\hat{\bar{Y}}_E}{\sqrt{g}} \dot{\sim} t_{(n')} \qquad (2.40)$$

if it can be shown that $(\bar{Y} - \hat{\bar{Y}}_E)$ and $n'g/V(\bar{Y} - \hat{\bar{Y}}_E)$ are independent.

To show this, notice that

$$(\bar{Y} - \hat{\bar{Y}}_E) = \frac{1}{MN}\{(mn - MN)\bar{y} + \sum_i^n \sum_j^{M-m} y_{ij} + \sum_i^{N-n} \sum_j^{M} y_{ij}\} \tag{2.41}$$

where $\sum_i^n \sum_j^{M-m} y_{ij}$ represents the sum of those non-sampled elements whose primaries were selected and $\sum_i^{N-n} \sum_j^{M} y_{ij}$ represents the sum of those non-sampled elements whose primaries were not selected.

Since g is a function of the sampled elements, and $\mathrm{Cov}(y_{ij}, y_{kl}) = 0$ for $i \neq k$, then g is independent of $\sum_i^{N-n} \sum_j^{M} y_{ij}$. To show $\sum_i^n \sum_j^{M-m} y_{ij}$ is independent of g, it must be shown to be independent of both $s_b^2$ and $s_w^2$. Defining

$$\bar{y}_r = \frac{\sum_j^m y_{rj}}{m}, \qquad r = 1, \ldots, n, \tag{2.42}$$

it can be seen that $\mathrm{Cov}(\sum_i^n \sum_j^{M-m} y_{ij}, \bar{y}_r - \bar{y}) = 0$. Since the $y_{ij}$ are normal random variables, $\sum_i^n \sum_j^{M-m} y_{ij}$ must be independent of any combination of $(\bar{y}_1 - \bar{y}, \ldots, \bar{y}_n - \bar{y})$, and therefore independent of $s_b^2$. Likewise, $\mathrm{Cov}(\sum_i^n \sum_j^{M-m} y_{ij}, \bar{y}_{rj} - \bar{y}_r) = 0$ and $\sum_i^n \sum_j^{M-m} y_{ij}$ is independent of $s_w^2$. Finally, it must be shown that $\bar{y}$ is independent of both $s_b^2$ and $s_w^2$. Observe that $\mathrm{Cov}(\bar{y}_i, \bar{y}_u) = 0$ for $i \neq u$, and the $\bar{y}_i$ are therefore distributed as normal random variables with mean $\mu$ and variance $\sigma_\alpha^2 + \frac{\sigma_\varepsilon^2}{m}$. A well known result is that $\bar{y}$ is stochastically independent of $(\bar{y}_1 - \bar{y}, \ldots, \bar{y}_n - \bar{y})$ (see, e.g., Hogg and Craig [1970, p. 163]). Hence, $\bar{y}$ is independent of any combination of $(\bar{y}_1 - \bar{y}, \ldots, \bar{y}_n - \bar{y})$, and therefore independent of $s_b^2$. Now observe that $\mathrm{Cov}(y_{ij} - \bar{y}_i, \bar{y}_i) = 0$ and hence $(y_{ij} - \bar{y}_i)$ is independent of $\bar{y}_i$. Therefore, $\bar{y} = \sum_i^n \bar{y}_i/n$ is independent of $(y_{ij} - \bar{y}_i)$ and likewise $s_w^2$.

Since $(\bar{Y} - \hat{\bar{Y}}_E)$ and $n'g/V(\bar{Y} - \hat{\bar{Y}}_E)$ are independent, an approximate $100(1-\alpha)\%$ confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}}_E \pm t_{\alpha/2;n'} \sqrt{g}\,] \;. \tag{2.43}$$

Due to the particular form of $V(\bar{Y} - \hat{\bar{Y}}_E)$, an exact confidence interval on $\bar{Y}$ not previously discussed in the literature can by developed by considering contrasts of $y_{ij}$. Let

$$u_i = c_1\bar{y}_i + c_2 d_i \tag{2.44}$$

where

$$\bar{y}_i = \frac{\sum_j^m y_{ij}}{m} \;, \tag{2.45}$$

$$d_i = \sum_j^m \ell_{ij} y_{ij} \;, \tag{2.46}$$

$$\sum_j^m \ell_{ij} = 0 \;, \tag{2.47}$$

$$c_3 = \sum_j^m \ell_{ij}^2 \;, \tag{2.48}$$

and the $\ell_{ij}$'s, $c_1$, $c_2$, and $c_3$ are constants. Assuming the model in (2.1) - (2.3),

$$d_i \sim N(0, \sigma_\varepsilon^2 c_3) \ , \tag{2.49}$$

$$\bar{y}_i \sim N(\mu, \sigma_\alpha^2 + \frac{1}{m} \sigma_\varepsilon^2) \ , \tag{2.50}$$

and

$$u_i \sim N(c_1\mu, c_1^2(\sigma_\alpha^2 + \frac{1}{m} \sigma_\varepsilon^2) + c_2^2 c_3 \sigma_\varepsilon^2) \tag{2.51}$$

as $\bar{y}_i$ and $d_i$ are independent. The independence of $\bar{y}_i$ and $d_i$ can be shown by the following argument. Since $E(d_i) = 0$,

$$\begin{aligned} \mathrm{Cov}(d_i, \bar{y}_i) &= E(d_i \bar{y}_i) \\ &= \frac{1}{m}(\sum_j^m \ell_{ij} E(y_{ij}^2) + \sum_j^m \sum_{j' \neq j}^m \ell_{ij} E(y_{ij} \ y_{ij'})) \ . \end{aligned} \tag{2.52}$$

Now

$$E(y_{ij}^2) = \sigma_\alpha^2 + \sigma_\varepsilon^2 + \mu^2 \tag{2.53}$$

and

$$E(y_{ij} \ y_{ij'}) = \sigma_\alpha^2 + \mu^2 \tag{2.54}$$

are both constants with respect to j and j'. Therefore, since $\sum_j^m \ell_{ij} = 0$, it follows that $\mathrm{Cov}(d_i, \bar{y}_i) = 0$. Since $d_i$ and $\bar{y}_i$ are normal random variables they are therefore independent.

The problem now is to choose $c_1$, $c_2$, and $c_3$ in such a manner that $V(u_i) = V(\bar{Y} - \hat{\bar{Y}}_E)$. Equating $V(u_i)$ to $V(\bar{Y} - \hat{\bar{Y}}_E)$ and solving for $c_1^2$ and $c_2^2 c_3$ yields

$$c_1^2 = \frac{1}{n}\left(1 - \frac{n}{N}\right) \tag{2.55}$$

and

$$c_2^2 c_3 = \frac{1}{mN}\left(1 - \frac{m}{M}\right) . \tag{2.56}$$

Of course, if $\bar{u} = \sum_i^n u_i/n$, then

$$\frac{\sum_i^n (u_i - \bar{u})^2}{V(u_i)} \sim \chi^2_{(n-1)} \tag{2.57}$$

(see, e.g., Hogg and Craig [1970, p. 165]). Now with $c_1^2 = \frac{1}{n}(1 - \frac{n}{N})$ and $c_2^2 c_3 = \frac{1}{mN}(1 - \frac{m}{M})$, then $V(u_i) = V(\bar{Y} - \hat{\bar{Y}}_E)$, and

$$\frac{(n-1)g_e}{V(\bar{Y} - \hat{\bar{Y}}_E)} \sim \chi^2_{(n-1)} \tag{2.58}$$

where

$$g_e = \frac{\sum_i^n (u_i - \bar{u})^2}{n-1} \tag{2.59}$$

and

$$u_i = \sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)}\,\bar{y}_i + \sqrt{\frac{1}{mN}\left(1 - \frac{m}{M}\right)}\;\frac{\sum_j^m \ell_{ij} y_{ij}}{\sqrt{\sum_j^m \ell_{ij}^2}} \; . \tag{2.60}$$

Hence, if it can be shown that $(\bar{Y} - \hat{\bar{Y}}_E)$ is independent of $(n-1)g_e/V(\bar{Y} - \hat{\bar{Y}}_E)$, then

$$\frac{\bar{Y}-\hat{\bar{Y}}_E}{\sqrt{g_e}} \sim t_{(n-1)} \ . \qquad (2.61)$$

As before,

$$\bar{Y} - \hat{\bar{Y}}_E = \frac{1}{MN}\{(mn - MN)\bar{y} + \sum_i^n \sum_j^{M-m} y_{ij} + \sum_i^{N-n} \sum_j^{M} y_{ij}\} \qquad (2.62)$$

and $\sum_i^{N-n} \sum_j^{M} y_{ij}$ is independent of $g_e$ which is a function of sample elements only. Now,

$$u_i - \bar{u} = c_1(\bar{y}_i - \bar{y}) + c_2\left(\sum_j^m \ell_{ij}y_{ij} - \frac{\sum\sum_{ij}^{nm} \ell_{ij}y_{ij}}{n}\right) \qquad (2.63)$$

and using the results stated earlier in this section, $\sum_i^n \sum_j^{M-m} y_{ij}$ and $\bar{y}$ are both independent of $(\bar{y}_i - \bar{y})$ and $\sum_j^m \ell_{ij}(y_{ij} - \bar{y}_i) = \sum_j^m \ell_{ij}y_{ij}$. An exact $100(1-\alpha)\%$ confidence interval on $\bar{Y}$ is therefore

$$[\hat{\bar{Y}}_E \pm t_{\alpha/2;n-1} \sqrt{g_e}\,] \ . \qquad (2.64)$$

A numerical example of this procedure is given in Appendix A.

It should be noted that this is an exact confidence interval for any choice of $\ell_{ij}$ as long as $\sum_j^m \ell_{ij} = 0$ and $\sum_j^m \ell_{ij}^2 = c_3$ for all i. Since the length of the confidence interval is determined by the value of $\sqrt{g_e}$, it would seem to be important to minimize this quantity when selecting the $\ell_{ij}$ and the corresponding value for $c_3$. However, since the distribution of $g_e$ does not depend on $\ell_{ij}$, any convenient set of $\ell_{ij}$ may be used. For example, if m is even, let

$$\begin{aligned} \ell_{ij} &= -1\ , \quad j = 1, \ldots, \frac{m}{2}\ , \\ &= +1\ , \quad j = \frac{m}{2} + 1, \ldots, m\ , \end{aligned} \tag{2.65}$$

and, if m is odd, let

$$\begin{aligned} \ell_{ij} &= -1\ , \quad j = 1, \ldots, \frac{m-1}{2}\ , \\ &= 0\ , \quad j = \frac{m+1}{2}\ , \\ &= +1\ , \quad j = \frac{m+3}{2}\ , \ldots, m\ , \end{aligned} \tag{2.66}$$

for all i. The robustness of the confidence interval to model breakdown may however depend on the $\ell_{ij}$, and this problem is discussed in Section 2.5.

## 2.3 Comparison of Confidence Intervals

The confidence intervals in (2.43) and (2.64) are now compared. Of course if $\rho$ is known, (2.19) provides an exact confidence interval with mn-1 degrees of freedom and would be superior to either one. Disregarding any consideration of the "goodness" of the approximation in (2.43), the criterion for comparison is the degrees of freedom associated with the t-statistic.

The degrees of freedom in the exact t-statistic of (2.64) are n-1 and in (2.43) they are n' where n' is defined in (2.37). After some algebraic simplifications,

$$n' = (n - 1)K \tag{2.67}$$

where

$$K = \frac{(1+d_1\delta)^2}{(1+d_2\delta^2)} , \tag{2.68}$$

$$\delta = \frac{\sigma_2^2}{\sigma_1^2} , \tag{2.69}$$

$$d_1 = \frac{p}{(1-p)} (1-q) , \tag{2.70}$$

$$d_2 = \frac{Np-1}{Np(Mq-1)} , \tag{2.71}$$

$$p = \frac{n}{N} , \tag{2.72}$$

and

$$q = \frac{m}{M} . \tag{2.73}$$

Whenever K is greater than one, the expected length of the approximate confidence interval will be less than the expected length of the exact confidence interval. Notice that K is always greater than one for $\delta \varepsilon [0, 1]$. This can be seen by noting that $K(0) = 1$, $K(1) > 1$, and that the only possible inflection point for $\delta$ in [0, 1] is a maximum i.e., the only possible $\delta$ in [0, 1] such that $K'(\delta) = 0$ also has $K''(\delta) < 0$. Table 2 gives the values of K for selected values of $\delta$, p, and q, in a population with $N = 20$ and $M = 100$.

TABLE 2

Values of K with N = 20 and M = 100

| | | | | $\delta$ | |
|---|---|---|---|---|---|
| p | q | .05 | .10 | .50 | 1.00 |
| .2 | .2 | 1.020 | 1.040 | 1.210 | 1.438 |
| .2 | .4 | 1.015 | 1.030 | 1.155 | 1.322 |
| .2 | .6 | 1.010 | 1.020 | 1.102 | 1.210 |
| .2 | .8 | 1.005 | 1.010 | 1.051 | 1.102 |
| .4 | .2 | 1.054 | 1.109 | 1.599 | 2.321 |
| .4 | .4 | 1.040 | 1.082 | 1.439 | 1.953 |
| .4 | .6 | 1.027 | 1.054 | 1.284 | 1.603 |
| .4 | .8 | 1.013 | 1.027 | 1.138 | 1.284 |
| .6 | .2 | 1.123 | 1.254 | 2.516 | 4.526 |
| .6 | .4 | 1.092 | 1.188 | 2.093 | 3.543 |
| .6 | .6 | 1.061 | 1.124 | 1.688 | 2.546 |
| .6 | .8 | 1.030 | 1.061 | 1.322 | 1.688 |
| .8 | .2 | 1.344 | 1.734 | 6.002 | 11.719 |
| .8 | .4 | 1.254 | 1.535 | 4.678 | 10.154 |
| .8 | .6 | 1.166 | 1.345 | 3.207 | 6.496 |
| .8 | .8 | 1.082 | 1.166 | 1.956 | 3.216 |

Although the degrees of freedom are less for the exact confidence interval than for the approximate interval, when $\delta$ is small, i.e., when the between primary variation is larger than the within primary variation, K is close to one and the degrees of freedom are nearly the same. In most survey populations, $\delta$ is in fact small, and use of the exact confidence interval seems advantageous. Even when $\delta$ and K are large, the t-values associated with the two intervals will not vary greatly if n is large, and the exact interval again seems appropriate.

## 2.4 Estimation of the Finite Population Total

The corresponding results for the population total $Y = \sum\sum_{ij}^{NM} y_{ij}$ follow immediately. Selecting the estimator

$$\hat{Y}_E = \frac{MN}{mn} \sum\sum_{ij}^{nm} y_{ij} \tag{2.74}$$

implies $E(Y - \hat{Y}_E) = 0$ and

$$V(Y - \hat{Y}_E) = M^2N^2 \{ \frac{\sigma_\alpha^2}{n}(1 - \frac{n}{N}) + \frac{\sigma_\varepsilon^2}{mn}(1 - \frac{mn}{MN})\} . \tag{2.75}$$

The unbiased estimator of $V(Y - \hat{Y}_E)$ is

$$g = \frac{M^2N^2}{mn}(1 - \frac{n}{N})s_b^2 + \frac{M^2N}{m}(1 - \frac{m}{M})s_w^2 . \tag{2.76}$$

The approximate $100(1-\alpha)\%$ confidence interval on Y is

$$[\hat{Y}_E \pm t_{\alpha/2;n'} \sqrt{g}\,] \qquad (2.77)$$

where g is as defined in (2.76). The exact 100(1-$\alpha$)% confidence interval on Y is

$$[\hat{Y}_E \pm t_{\alpha/2;n-1}\sqrt{g_e}\,] \qquad (2.78)$$

where $g_e$ is identical to (2.59) except that

$$c_1^2 = \frac{M^2N^2}{n}\left(\frac{N-n}{N}\right) \qquad (2.79)$$

and

$$c_2^2c_3 = \frac{MN}{m}(M - m) \ . \qquad (2.80)$$

## 2.5 Robustness to Model Breakdown

The results of Section 2.1 and Section 2.2 were developed under the assumptions stated in (2.1) - (2.3). This section examines the robustness of $\hat{\hat{Y}}_E$ and the exact confidence interval to a breakdown in the assumed model. Notice that if primaries are drawn without replacement and secondaries are sampled independently within each primary, $\hat{\hat{Y}}_E$ is the classical unbiased estimator for $\bar{Y}$. Therefore, since

$$E(\bar{Y} - \hat{\hat{Y}}_E \mid \text{any finite population}) = 0 \ , \qquad (2.81)$$

it follows that $E(\bar{Y} - \hat{\hat{Y}}_E) = 0$ no matter what is assumed about the super-population.

Similarly, the estimator g in (2.35) is the classical unbiased estimator for $V(\hat{\bar{Y}}_E)$ if primaries are selected without replacement and secondaries are randomly selected within each primary. Hence, $E(g - V(\bar{Y} - \hat{\bar{Y}}_E)) = 0$ for any assumed super-population model.

The robustness of the exact confidence interval in (2.64) to model breakdown must rely on the robustness of the t-statistic. However, for a specific model breakdown, it may be possible to select the $\ell_{ij}$'s so that the consequences are minimized.

As an example, consider the situation where the $\varepsilon_{ij}$'s have a non-normal distribution with

$$E(\varepsilon_{ij}) = 0 \ , \qquad (2.82)$$

$$E(\varepsilon_{ij}^2) = \sigma_\varepsilon^2 \ , \qquad (2.83)$$

$$E(\varepsilon_{ij}^3) = \theta \ , \qquad (2.84)$$

and

$$E(\varepsilon_{ij}^4) = \omega \ . \qquad (2.85)$$

Although

$$u_i = c_1\bar{y}_i + c_2 d_i \qquad (2.86)$$

is now non-normal, an appropriate choice of the $\ell_{ij}$'s can minimize the non-normality. Since the $\ell_{ij}$'s affect $u_i$ only through the variable $d_i$, they should be chosen so that $d_i$ behaves as a normal random variable. For the $i^{th}$ primary, let

$$\xi_j = \ell_{ij}\varepsilon_{ij} \qquad (2.87)$$

and

$$d_i = \sum_j^m \xi_j \ . \tag{2.88}$$

From (2.82) and (2.83), it follows that

$$E(\xi_j) = 0 \tag{2.89}$$

and

$$V(\xi_j) = \ell_{ij}^2 \sigma_\varepsilon^2 \ . \tag{2.90}$$

Notice that if $|\ell_{ij}|$ is equal for all j, the $\xi_j$'s are independent random variables with equal first and second moments. The $\xi_j$'s then satisfy the Lindeberg condition, and $d_i$ has a limiting distribution which is normal (see, e.g., Gnedenko [1963, p. 290]). Hence, if the $\ell_{ij}$'s are selected such that $|\ell_{ij}|$ is equal for all j, the non-normality of $u_i$ is reduced.

It also seems desirable to choose the $\ell_{ij}$'s so that the third and fourth central moments of $d_i$ correspond to those of a normal random variable. This implies setting

$$E(d_i^3) = 0 \tag{2.91}$$

and

$$E(d_i^4) = 3(E(d_i^2))^2 \ . \tag{2.92}$$

Keeping $\sum_j^m \ell_{ij} = 0$, $\sum_j^m \ell_{ij}^2 = c_3$, and using (2.82) - (2.85), it follows that

$$E(d_i) = \sum_j^m \ell_{ij} E(\varepsilon_{ij}) = 0 \ , \tag{2.93}$$

$$E(d_i^2) = \sigma_\varepsilon^2 c_3 \ , \tag{2.94}$$

$$E(d_i^3) = \theta \sum_j^m \ell_{ij}^3 \ , \tag{2.95}$$

and

$$E(d_i^4) = \omega \sum_j^m \ell_{ij}^4 + 6\ \sigma_\varepsilon^4 \sum_{j<j'}^m \ell_{ij}^2 \ell_{ij'}^2 \ . \tag{2.96}$$

Therefore, satisfying (2.91) and (2.92) implies

$$\sum_j^m \ell_{ij}^3 = 0 \tag{2.97}$$

and

$$\sum_j^m \ell_{ij}^4 \left( \frac{\omega}{3\ \sigma_\varepsilon^4} - 1 \right) = 0 \ . \tag{2.98}$$

If $|\ell_{ij}|$ is equal for all j, (2.97) is satisfied, and for $|\ell_{ij}| \neq 0$, (2.98) is most closely satisfied with small values of $\ell_{ij}$. Hence, to reduce the non-normality of $u_i$ when $\varepsilon_{ij}$ is non-normal, select the $\ell_{ij}$'s so that $|\ell_{ij}|$ is equal for all j, and $\sum_j^m \ell_{ij}^4$ is small.

As another example, assume that the $\varepsilon_{ij}$ are normally distributed, but that $V(\varepsilon_{ij}) = \sigma_i^2$, where $\sigma_i^2$ is not the same for all i. Under such a model,

$$V(u_i) = c_1^2 \sigma_\alpha^2 + \sigma_i^2 (\frac{c_1^2}{m} + c_2^2 c_3) \tag{2.99}$$

is no longer equal for all i. However, if the values of $\sigma_i^2$ are known or can be estimated, different values of $c_2^2c_3$ can be chosen for each primary so that $\sigma_i^2(\frac{c_1^2}{m} + (c_2^2c_3)_i)$ will be equal for all primaries.

Since no choice of the $\ell_{ij}$'s is best for every possible model breakdown, the $\ell_{ij}$'s should be chosen to protect against the breakdown that is most likely to occur in a particular situation.

# 3. TWO-STAGE SAMPLING WITH PRIMARIES OF UNEQUAL SIZE

## 3.1 Definition of the Model

A two-stage sampling design will now be considered for the case when primaries are not of equal size. Let

$N$ = number of primaries in finite population,

$n$ = number of primaries sampled,

$M_i$ = size of $i^{th}$ primary,

$m_i$ = number of secondaries selected from the $i^{th}$ primary,

$M_o = \sum_i^N M_i$ = total elements in population, and

$m_o = \sum_i^n m_i$ = total elements in sample.

It is assumed that all of these variables are known. The superpopulation is again represented as

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{3.1}$$

where

$$\alpha_i \sim N(0, \sigma_\alpha^2) , \tag{3.2}$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) , \tag{3.3}$$

and $\alpha_i$ and $\varepsilon_{ij}$ are independent. As in Section 2.1, redefine (3.1) as

$$y_{ij} = \mu + \eta_{ij} \tag{3.4}$$

where the $\eta_{ij}$'s are independent normal random variables with mean zero and

$$
\begin{aligned}
E(\eta_{ij}\ \eta_{k\ell}) &= \sigma_\alpha^2 + \sigma_\varepsilon^2\ , && i=k,\ j=\ell\ , \\
&= \sigma_\alpha^2\ , && i=k,\ j\neq\ell\ , \\
&= 0\ , && i\neq k\ . \qquad (3.5)
\end{aligned}
$$

The finite population is represented as

$$\underline{Y} = \underline{1}\mu + \underline{H} \qquad (3.6)$$

where $\underline{Y}$ is the $(M_o\times 1)$ vector of finite population observations, $\underline{1}$ is a $(M_o\times 1)$ vector of ones, and $\underline{H}$ is a $(M_o\times 1)$ vector of random variables. The covariance matrix for $\underline{H}$ is now the $(M_o\times M_o)$ block diagonal matrix

$$\sigma_\varepsilon^2 V = \sigma_\varepsilon^2 \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \dots & V_N \end{bmatrix} \qquad (3.7)$$

where

$$V_i = I_{M_i} + \rho J_{M_i}\ , \qquad i = 1, \dots, N\ , \qquad (3.8)$$

with $I_{M_i}$ denoting the identity matrix of order $M_i$, $J_{M_i}$ denoting the $(M_i\times M_i)$ matrix with all elements equal to one, and $\rho = \sigma_\alpha^2/\sigma_\varepsilon^2$.

Recall that in Section 2.1, the finite population mean was expressible as the BLUE of $\mu$. However, in the present situation where primaries are of unequal size, the BLUE of $\mu$ is

$$\hat{\mu} = (\underline{1}^T V^{-1} \underline{1})^{-1} \underline{1}^T V^{-1} \underline{Y}$$

$$= \frac{\sum_{i}^{N} \sum_{j}^{M_i} y_{ij} \left(\frac{1}{1+M_i\rho}\right)}{\sum_{i}^{N} M_i \left(\frac{1}{1+M_i\rho}\right)} \quad . \tag{3.9}$$

Since there is no linear transformation of $\hat{\mu}$ that equals $\bar{Y}$, $\hat{\mu}$ does not now readily suggest an estimator of $\bar{Y}$. For this reason, a different method for the estimation of $\bar{Y}$ is introduced.

## 3.2 Estimation and Variance Formulas for the Finite Population Mean

Express the finite population in terms of the least squares fit

$$y_{ij} = a_i + e_{ij} \,, \qquad i = 1, \ldots, N \,, \quad j = 1, \ldots, M_i \,, \tag{3.10}$$

where

$$a_i = \frac{\sum_{j}^{M_i} y_{ij}}{M_i} \tag{3.11}$$

minimizes $\sum_{i}^{N} \sum_{j}^{M_i} (y_{ij} - a_i)^2$. The finite population mean is

$$\bar{Y} = \frac{\sum_{i}^{N} M_i \bar{Y}_i}{M_o}$$

$$= \frac{\sum_{i}^{N} M_i a_i}{M_o} . \qquad (3.12)$$

Obviously, an estimator of $\bar{Y}$ such that $E(\bar{Y} - \hat{\bar{Y}}) = 0$ is

$$\hat{\bar{Y}}_U = \sum_{i}^{N} \frac{M_i}{M_o} \hat{a}_i \qquad (3.13)$$

where $E(a_i - \hat{a}_i) = 0$.

For the primaries selected in the sample,

$$\hat{a}_i = \frac{\sum_{j}^{m_i} y_{ij}}{m_i} = \bar{y}_i \qquad (3.14)$$

is the classical estimator of $a_i$. For the primaries of the finite population not represented in the sample, the only knowledge about them is that the $a_i$ have been selected from a population with mean $\mu$. A logical estimator for $a_i$ is therefore the BLUE of $\mu$. The BLUE of $\mu$ computed from the sample is

$$\hat{\mu}_s = \frac{\sum_{i}^{n} \bar{y}_i \left(\frac{1}{V(\bar{y}_i)}\right)}{\sum_{i}^{n} \left(\frac{1}{V(\bar{y}_i)}\right)} \qquad (3.15)$$

where

$$V(\bar{y}_i) = \sigma_\alpha^2 + \sigma_\varepsilon^2/m_i \ . \tag{3.16}$$

Unfortunately, $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$ are seldom known in practice and $V(\bar{y}_i)$ cannot be computed. Koch [1967] has considered three other unbiased estimators of $\mu$ for the model assumed in (3.1) - (3.3). These three estimators are

$$\bar{y}_a = \sum_i^n k_i m_i \bar{y}_i \ , \tag{3.17}$$

$$\bar{y}_b = \frac{\sum_i^n m_i \bar{y}_i}{m_o} \ , \tag{3.18}$$

and

$$\bar{y}_c = \frac{\sum_i^n \bar{y}_i}{n} \tag{3.19}$$

where

$$k_i = \frac{m_o - m_i}{m_o^2 - \sum_i^n m_i^2} \ . \tag{3.20}$$

Of course when all $m_i$ are equal, all three estimators are the same.

To compare the three estimators it is appropriate to examine their variances. It can be seen that

$$V(\bar{y}_a) = (\sum_i^n k_i^2 m_i^2)\sigma_\alpha^2 + (\sum_i^n k_i^2 m_i)\sigma_\varepsilon^2$$

$$= A_1\sigma_\alpha^2 + A_2\sigma_\varepsilon^2 , \quad (3.21)$$

$$V(\bar{y}_b) = \left(\frac{\sum_i^n m_i^2}{m_o^2}\right)\sigma_\alpha^2 + (\frac{1}{m_o})\sigma_\varepsilon^2$$

$$= B_1\sigma_\alpha^2 + B_2\sigma_\varepsilon^2 , \quad (3.22)$$

and

$$V(\bar{y}_c) = (\frac{1}{n})\sigma_\alpha^2 + \left(\frac{1}{n^2} \sum_i^n \frac{1}{m_i}\right)\sigma_\varepsilon^2$$

$$= C_1\sigma_\alpha^2 + C_2\sigma_\varepsilon^2 . \quad (3.23)$$

Koch presents a table with n = 5 and for various values of $m_i$ (i = 1, ..., 5), shows the following inequalities hold:

$$C_1 \leq A_1 \leq B_1 \quad (3.24)$$

for the coefficients of $\sigma_\alpha^2$, and

$$B_2 \leq A_2 \leq C_2 \quad (3.25)$$

for the coefficients of $\sigma_\varepsilon^2$. The solution to which of the three estimators has the smallest variance is therefore dependent on the actual values of $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$. Since $\bar{y}_a$ is the intermediate estimator in regard to the coefficients of $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$, it will be chosen as the estimator of $\mu$ for purposes of illustration. Koch further shows that $\bar{y}_a$ is optimal with respect to a particular quadratic, location-sensitive criterion.

Substituting (3.14) and (3.17) into (3.13) gives

$$\hat{\bar{Y}}_U = \pi\bar{y}_r + (1 - \pi)\bar{y}_a \tag{3.26}$$

where

$$\pi = \frac{\sum_i^n M_i}{M_o} \tag{3.27}$$

and

$$\bar{y}_r = \frac{\sum_i^n M_i\bar{y}_i}{\sum_i^n M_i} \; . \tag{3.28}$$

Notice that $\bar{y}_r$ is the classical ratio estimator of $\bar{Y}$ for two-stage sampling when units are selected with equal probabilities (see, e.g., Cochran [1963, p. 300]).

The variance of $(\bar{Y} - \hat{\bar{Y}}_U)$ is computed by writing

$$\bar{Y} - \hat{\bar{Y}}_U = \frac{1}{M_o}\left\{\sum_i^{N-n} Y_i + \sum_i^n [Y_i - y_i(\frac{M_i}{m_i} + M_o(1 - \pi)k_i)]\right\} \tag{3.29}$$

where

$$Y_i = \sum_j^{M_i} y_{ij} \; , \tag{3.30}$$

$$y_i = \sum_j^{m_i} y_{ij} \,, \tag{3.31}$$

and $\sum_i^{N-n}$ denotes summation over the non-sampled primaries. Notice that

$$E(\bar{Y} - \hat{\bar{Y}}_U) = 0 \,, \tag{3.32}$$

$$V(Y_i) = M_i^2 \sigma_\alpha^2 + M_i \sigma_\varepsilon^2 \,, \tag{3.33}$$

$$V(y_i) = m_i^2 \sigma_\alpha^2 + m_i \sigma_\varepsilon^2 \,, \tag{3.34}$$

and

$$\mathrm{Cov}(Y_i,\, y_i) = m_i M_i \sigma_\alpha^2 + m_i \sigma_\varepsilon^2 \,. \tag{3.35}$$

Hence,

$$V(\bar{Y} - \hat{\bar{Y}}_U) = \frac{\sigma_\alpha^2}{M_o^2} \left\{ \sum_i^{N-n} M_i^2 + M_o^2 (1 - \pi)^2 \sum_i^n k_i^2 m_i^2 \right\}$$

$$+ \frac{\sigma_\varepsilon^2}{M_o^2} \left\{ \sum_i^n \frac{M_i^2}{m_i} + M_o (1 - 2\pi + (1 - \pi) [M_o (1 - \pi) \sum_i^n k_i^2 m_i \right.$$

$$\left. + 2 \sum_i^n k_i (M_i - m_i)]) \right\} \,. \tag{3.36}$$

If all $m_i = m$, this reduces to

$$V(\bar{Y} - \hat{\bar{Y}}_U) = \frac{\sigma_\alpha^2}{n}\left\{\frac{n \sum_{i}^{N-n} M_i^2}{M_o^2} + (1 - \pi)^2\right\}$$

$$+ \frac{\sigma_\varepsilon^2}{mn}\left\{\frac{n}{M_o}\left(\frac{\sum_{i}^{n} M_i^2}{M_o} - m\right) + (1 - \pi^2)\right\} . \qquad (3.37)$$

Notice if all $M_i = M$, (3.37) reduces to

$$V(\bar{Y} - \hat{\bar{Y}}_U) = \frac{\sigma_\alpha^2}{n}(1 - \frac{n}{N}) + \frac{\sigma_\varepsilon^2}{mn}(1 - \frac{mn}{MN}) , \qquad (3.38)$$

which is the same result as (2.17) in Section 2.1.

A comment should be made about the preceding results and those to follow in section 4. When primaries are of unequal size, even though $M_1, \ldots, M_N$ are fixed and assumed known, in a strict probabilistic sense the $M_1, \ldots, M_n$ corresponding to the sampled primaries are really random variables whose realization depends upon which primaries are sampled. Hence, the arguments used above and those to follow in Section 4 are really conditional arguments for given values of $M_1, \ldots, M_n$. However, since the unbiasedness of the estimators of $\bar{Y}$ and the confidence levels of the corresponding confidence intervals will not depend upon the values of $M_1, \ldots, M_n$, these properties will also apply in an unconditional sense.

## 3.3 Confidence Intervals on $\bar{Y}$

The problem of estimating a linear combination of variance components such as $V(\bar{Y} - \hat{\bar{Y}}_U)$ is much more difficult for the unbalanced case, i.e., all $m_i$ not equal, than for the balanced case considered in Section 2.2. Searle [1971b] cites two problems in the unbalanced case; namely, several methods of estimation are available with no clear decision on which is best, and all methods involve cumbersome algebra. Searle [1971a, 1971b] gives an extensive survey of various methods to estimate variance components. Perhaps the most popular method is the analysis of variance method suggested by Henderson [1953], which for the model assumed in (3.1) - (3.3), yields the unbiased estimators

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_i^n \sum_j^{m_i} (y_{ij} - \bar{y}_i)^2}{m_o - n} \tag{3.39}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{\sum_i^n m_i(\bar{y}_i - \bar{y})^2 - (n-1)\hat{\sigma}_\varepsilon^2}{m_o - \frac{\sum_i^n m_i^2}{m_o}} . \tag{3.40}$$

An unbiased estimator for $V(\bar{Y} - \hat{\bar{Y}}_U)$ would therefore be

$$g_u = b_1\hat{\sigma}_\alpha^2 + b_2\hat{\sigma}_\varepsilon^2 \tag{3.41}$$

where

$$b_1 = \frac{1}{M_o^2}\left\{\sum_i^{N-n} M_i^2 + M_o^2(1 - \pi)^2 \sum_i^n k_i^2 m_i^2\right\} \tag{3.42}$$

and

$$b_2 = \frac{1}{M_o^2}\left\{\sum_i^n \frac{M_i^2}{m_i} + M_o(1 - 2\pi + (1 - \pi)[M_o(1 - \pi)\sum_i^n k_i^2 m_i + 2\sum_i^n k_i(M_i - m_i)])\right\} . \tag{3.43}$$

However, as Searle [1971a, 1971b] reports, the distributions of the variance component estimators are unknown, although they may be expressed as linear combinations of non-central chi-squares as discussed by Harville [1969]. The only partial exception is that under the assumption that the random effects have normal distributions,

$$\hat{\sigma}_\varepsilon^2 \sim \chi^2_{(m_o-n)}\left(\frac{\sigma_\varepsilon^2}{m_o-n}\right) . \tag{3.44}$$

Obviously since the distributions of the variance component estimators themselves are unknown, the distribution of a linear combination such as $g_u$ is also unknown.

One method to approximate the distribution of $g_u$ is to equate its first two moments to those of a chi-square variable. That is, let

$$g_u \mathrel{\dot\sim} b\chi^2_{(f)} , \tag{3.45}$$

$$E(g_u) = b_1\sigma_\alpha^2 + b_2\sigma_\varepsilon^2 \tag{3.46}$$

$$= bf ,$$

and

$$V(g_u) = b_1^2V(\hat{\sigma}_\alpha^2) + b_2^2V(\hat{\sigma}_\varepsilon^2) + 2b_1b_2 \text{ Cov}(\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2)$$

$$= 2b^2f . \tag{3.47}$$

Solving for b and f gives

$$b = \frac{b_1^2V(\hat{\sigma}_\alpha^2)+b_2^2V(\hat{\sigma}_\varepsilon^2)+2b_1b_2 \text{ Cov}(\hat{\sigma}_\alpha^2,\hat{\sigma}_\varepsilon^2)}{2(b_1\sigma_\alpha^2+b_2\sigma_\varepsilon^2)} \tag{3.48}$$

and

$$f = \frac{2(b_1^2\sigma_\alpha^4+b_2^2\sigma_\varepsilon^4+2b_1b_2\sigma_\alpha^2\sigma_\varepsilon^2)}{b_1^2V(\hat{\sigma}_\alpha^2)+b_2^2V(\hat{\sigma}_\varepsilon^2)+2b_1b_2 \text{ Cov}(\hat{\sigma}_\alpha^2,\hat{\sigma}_\varepsilon^2)} . \tag{3.49}$$

For the model assumed in (3.1) - (3.3) Searle [1971a] gives formulas for $V(\hat{\sigma}_\alpha^2)$, $V(\hat{\sigma}_\varepsilon^2)$, and $\text{Cov}(\hat{\sigma}_\alpha^2, \hat{\sigma}_\varepsilon^2)$ under the normality assumptions. However, since these terms are functions of squares and products of $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$, he further cites results derived by Ahrens which provide unbiased estimators for them. At best the above procedure would lead to only an approximate distribution of $g_u$, and since estimators of unknown parameters are used when solving for b and f, the method appears questionable.

Since the distribution of $g_u$ is not available for the unbalanced case, it appears the simplifying assumption that all $m_i = m$ must be made to develop any useful results. This assumption is not unrealistic, since in many surveys, the desire for an equal surveyor workload requires the selection of an equal number of secondaries for each sampled primary. By assuming equal secondary sampling, the results developed in Section 2.2 may now be used to construct confidence intervals on $\bar{Y}$. When all $m_i = m$,

$$V(\bar{Y} - \hat{\bar{Y}}_U) = g_1\sigma_1^2 + g_2\sigma_2^2 \tag{3.50}$$

where

$$g_1 = \frac{\sum_i^{N-n} M_i^2}{M_o^2 m} + \frac{(1-\pi)^2}{mn} , \tag{3.51}$$

$$g_2 = \frac{(\sum_i^{n} M_i^2 - \sum_i^{N-n} M_i^2)}{M_o^2 m} - \frac{1}{M_o} + \frac{2(\pi)(1-\pi)}{mn} , \tag{3.52}$$

$$\sigma_1^2 = m\sigma_\alpha^2 + \sigma_\varepsilon^2 , \tag{3.53}$$

and

$$\sigma_2^2 = \sigma_\varepsilon^2 . \tag{3.54}$$

Using Theorem 2.1 of Section 2.2 gives

$$g = g_1 s_b^2 + g_2 s_w^2 \tag{3.55}$$

where

$$s_b^2 = \frac{\sum_i^n m(\bar{y}_i - \bar{y})^2}{n-1} \tag{3.56}$$

and

$$s_w^2 = \frac{\sum_{i}^{n}\sum_{j}^{m}(y_{ij}-\bar{y}_i)^2}{n(m-1)} \tag{3.57}$$

as an unbiased estimator of $V(\bar{Y} - \hat{\bar{Y}}_U)$. Furthermore,

$$\frac{n'g}{V(\bar{Y}-\hat{\bar{Y}}_U)} \overset{\bullet}{\sim} \chi^2_{(n')} \tag{3.58}$$

and

$$\frac{\bar{Y}-\hat{\bar{Y}}_U}{\sqrt{g}} \overset{\bullet}{\sim} t_{(n')} \tag{3.59}$$

where

$$n' = \frac{(g_1\sigma_1^2+g_2\sigma_2^2)^2}{\frac{g_1^2\sigma_1^4}{n_1} + \frac{g_2^2\sigma_2^4}{n_2}} , \tag{3.60}$$

$$n_1 = n - 1 , \tag{3.61}$$

and

$$n_2 = n(m - 1) . \tag{3.62}$$

An approximate $100(1-\alpha)\%$ confidence interval on $\bar{Y}$ is therefore given by

$$[\hat{\bar{Y}}_U \pm t_{\alpha/2;n'} \sqrt{g}\ ] \tag{3.63}$$

with g as defined in (3.55).

An exact confidence interval on $\bar{Y}$ may also be constructed. Consider the random variable

$$u_i = c_1\bar{y}_i + c_{2i}d_i \qquad (3.64)$$

where

$$d_i = \sum_j^{m_i} \ell_{ij}y_{ij} \ , \qquad (3.65)$$

$$\sum_j^{m_i} \ell_{ij} = 0 \ , \qquad i = 1, \ldots, n \ , \qquad (3.66)$$

$$\sum_j^{m_i} \ell_{ij}^2 = c_{3i}, \qquad i = 1, \ldots, n \ , \qquad (3.67)$$

and the $\ell_{ij}$'s, $c_1$, $c_{2i}$'s and $c_{3i}$'s are constants. Using the model stated in (3.1) - (3.3), it can be shown that

$$u_i \sim N(c_1\mu, c_1^2\ \sigma_\alpha^2 + (\frac{c_1^2}{m_i} + c_{2i}^2c_{3i})\sigma_\varepsilon^2 \qquad (3.68)$$

Equating $V(u_i)$ to $V(\bar{Y} - \hat{\bar{Y}}_U)$ in (3.36) and solving for $c_1^2$ and $c_{2i}^2 c_{3i}$ yields

$$c_1^2 = \frac{1}{M_o^2} \{ \sum_i^{N-n} M_i^2 + M_o^2(1 - \pi)^2 \sum_i^n k_i^2 m_i^2 \} \qquad (3.71)$$

and

$$c_{2i}^2 c_{3i} = \frac{-c_1^2}{m_i} + \frac{1}{M_o^2} \{\sum_i^n \frac{M_i^2}{m_i} + M_o(1 - 2\pi + (1 - \pi)[M_o(1 - \pi)\cdot$$

$$\sum_i^n k_i^2 m_i + 2\sum_i^n k_i(M_i - m_i)])\}. \qquad (3.72)$$

Then

$$\frac{\bar{Y}-\hat{\bar{Y}}_U}{\sqrt{g_e}} \sim t_{(n^* - 1)} \qquad (3.73)$$

where

$$g_e = \frac{\sum_i^{n^*} (u_i - \bar{u})^2}{n^* - 1}, \qquad (3.74)$$

$n^*$ = number of sampled primaries with $c_{2i}^2 c_{3i} \geq 0$,

and an exact $100(1 - \alpha)\%$ confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}}_U \pm t_{\alpha/2;n^*-1}\sqrt{g_e}]. \qquad (3.75)$$

As in the equal primary case, any convenient set of $\ell_{ij}$ may be used as long as $\sum_j^{m_i} \ell_{ij} = 0$ and $c_{2i}^2 c_{3i} \geq 0$ and satisfies (3.72).

### 3.4 Robustness of $\hat{\bar{Y}}_U$ to Model Breakdown

The robustness of $\hat{\bar{Y}}_U$ to model breakdown is now examined. Recall

$$\hat{\bar{Y}}_U = \frac{n}{N}\,\bar{y}_u + (1 - \pi)\bar{y}_a \tag{3.76}$$

where

$$\bar{y}_u = \frac{N}{nM_o}\sum_i^n M_i\bar{y}_i \ , \tag{3.77}$$

$$\bar{y}_a = \sum_i^n k_i m_i \bar{y}_i \ , \tag{3.78}$$

$$k_i = \frac{m_o - m_i}{m_o^2 - \sum_i^n m_i^2} \ , \tag{3.79}$$

and

$$\pi = \frac{\sum_i^n M_i}{M_o} \ . \tag{3.80}$$

Let $S_1$ denote that primaries are selected with equal probabilities and without replacement. Let $S_2$ denote that primaries are drawn with probability proportional to size and with replacement. In both $S_1$ and $S_2$ it is assumed that the secondaries are sampled at random. Then for a given finite population,

$$E(\hat{\bar{Y}}_U|S_1) = \frac{n}{N}\,\bar{Y} + \frac{(1-\pi)n}{N}\sum_i^N k_i m_i \bar{Y}_i \tag{3.81}$$

and

$$E(\hat{\bar{Y}}_U|S_2) = \frac{n}{M_o^2}\sum_i^N M_i^2 Y_i + \frac{(1-\pi)n}{M_o}\sum_i^N k_i m_i M_i \bar{Y}_i \ . \tag{3.82}$$

Thus, for a given finite population $\hat{\bar{Y}}_U$ is a biased estimator of $\bar{Y}$ under either $S_1$ or $S_2$ even though

$$E(\bar{y}_u | S_1) = \bar{Y} \tag{3.83}$$

and, if all $m_i$ are equal,

$$E(\bar{y}_a | S_2) = \bar{Y} . \tag{3.84}$$

For the special case when only one primary is selected, i.e., n = 1,

$$\hat{\bar{Y}}_U = \bar{y}_i , \tag{3.85}$$

$$E(\hat{\bar{Y}}_U | S_2) = \bar{Y} , \tag{3.86}$$

and hence $E(\bar{Y} - \hat{\bar{Y}}_U) = 0$ for any assumed super-population. This special case occurs when a stratified multi-stage design is used and only one primary is chosen from each strata.

Another special case of interest is stratified sampling, i.e., when all primaries are sampled. In this case, $\pi = 1$, and if secondaries are sampled at random from each primary

$$\hat{\bar{Y}}_U = \frac{\sum_i^N M_i \bar{y}_i}{M_o} \tag{3.87}$$

is the classical unbiased estimator of $\bar{Y}$. Hence, the unbiasedness of $\hat{\bar{Y}}_U$ in stratified sampling is robust to model breakdown. This result was also obtained by Hartley and Sielken [1975] who further discuss the robustness of the confidence interval for stratified sampling.

The results of Section 2.5 concerning selection of the $\ell_{ij}$'s also apply to the unequal primary case.

### 3.5 Comparison of $\hat{\hat{Y}}_U$ with the BLUE

The estimator $\hat{\hat{Y}}_U$ was motivated by the least squares estimators of the finite population parameters. An alternative unbiased estimator $\hat{\hat{Y}}$ for $\bar{Y}$ can be constructed by minimizing $V(\bar{Y} - \hat{\hat{Y}})$ subject to the restriction that

$$E(\bar{Y} - \hat{\hat{Y}}) = 0 . \tag{3.88}$$

In particular, let $\hat{\hat{Y}} = \sum_i^n h_i \bar{y}_i$ where the $h_i$'s do not depend on the $y_{ij}$'s. Then

$$E(\bar{Y} - \hat{\hat{Y}}) = \mu(1 - \sum_i^n h_i) , \tag{3.89}$$

and the condition that

$$\sum_i^n h_i = 1 \tag{3.90}$$

is imposed to insure that $\hat{\hat{Y}}$ is unbiased. Minimizing $V(\bar{Y} - \hat{\hat{Y}})$ with respect to $h_i$ subject to (3.90) yields

$$h_i = m_i \left\{\frac{1}{M_o} + \frac{(b_i+k)\lambda_i}{m_i}\right\} \tag{3.91}$$

where

$$b_i = \frac{1}{M_o}(M_i - m_i) \ , \tag{3.92}$$

$$k = \left( 1 - \frac{\sum_i^n M_i \lambda_i}{M_o} - \frac{\sum_i^n m_i (1-\lambda_i)}{M_o} \right) / \sum_i^n \lambda_i \ , \tag{3.93}$$

and

$$\lambda_i = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2 / m_i) \ . \tag{3.94}$$

Hence, the estimator $\hat{\bar{Y}}_S = \sum_i^n h_i \bar{y}_i$ with the $h_i$ as defined in (3.91) is the BLUE of $\bar{Y}$. This estimator has been previously suggested by Scott and Smith [1969]. However $\hat{\bar{Y}}_S$ requires knowledge of $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$, and these values are seldom known in practice. Scott and Smith make the simplifying assumption that all $m_i = m$, and recommend estimating $\lambda$ by

$$\hat{\lambda} = \frac{r-1}{r} \ , \qquad r \geq 1 \ ,$$

$$= 0 \ , \qquad r < 1 \tag{3.95}$$

where

$$r = \frac{s_b^2}{s_w^2} \ . \tag{3.96}$$

Notice that in the special case when $M_i = M$ and $m_i = m$,

$$h_i = \frac{1}{n} \ , \tag{3.97}$$

and

$$\hat{\bar{Y}}_S = \hat{\bar{Y}}_E \ . \tag{3.98}$$

If $m_i = m$ but not all $M_i$ are equal, then

$$h_i = \frac{1}{n} - \frac{\lambda}{M_o}\left(\frac{\sum_i^n M_i}{n} - M_i\right) , \tag{3.99}$$

and

$$\hat{\bar{Y}}_S = \pi_S \bar{y}_r + (1 - \pi_S)\bar{y} \tag{3.100}$$

where

$$\pi_S = \frac{\lambda \sum_i^n M_i}{M_o} , \tag{3.101}$$

$$\bar{y}_r = \sum_i^n M_i \bar{y}_i / \sum_i^n M_i , \tag{3.102}$$

and

$$\bar{y} = \frac{\sum\sum_{ij}^{nm} y_{ij}}{mn} . \tag{3.103}$$

The estimator $\hat{\bar{Y}}_U$ in (3.26) is therefore a special case of $\hat{\bar{Y}}_S$ in which $\lambda = 1$.

Even though $\hat{\bar{Y}}_S$ minimizes $V(\bar{Y} - \hat{\bar{Y}})$, $\hat{\bar{Y}}_U$ has the advantage that under the assumed model, its distribution and the distribution of its estimated variance are known, and confidence intervals can therefore be constructed on $\bar{Y}$. On the other hand, Scott and Smith do not derive the distribution of $\hat{\bar{Y}}_S$ when $\lambda$ is unknown, and therefore make no comments concerning confidence intervals. Further, $\lambda$ will indeed be close to one if $\sigma_\varepsilon^2/m$ is small relative to $\sigma_\alpha^2$, as would be the case if there is a clustering effect within primaries. As stated in Section 2.3, this type of survey population is quite common. Finally, if m is large and the $M_i$ are of relatively equal size, the difference between $V(\bar{Y} - \hat{\bar{Y}}_S)$ and $V(\bar{Y} - \hat{\bar{Y}}_U)$ will be small. Hence, although $\hat{\bar{Y}}_S$ is the BLUE of $\bar{Y}$, the fact that its distributional properties are not known when $\lambda$ is unknown, suggests that $\hat{\bar{Y}}_U$ is of greater practical value.

# 4. REGRESSION ESTIMATORS IN TWO-STAGE SAMPLING

## 4.1 Definition of the Model

There are many situations in survey sampling in which the variable of interest, y, is related to another variable, x. If such a situation occurs, a more appropriate super-population model than the one assumed in Section 3.1 is

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij} \tag{4.1}$$

where

$$\alpha_i \sim N(0, \sigma_\alpha^2) \ , \tag{4.2}$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \ , \tag{4.3}$$

$\alpha_i$, $x_{ij}$, and $\varepsilon_{ij}$ are all independent, and $\mu$ and $\beta$ are constants.

An estimator for the finite population mean is developed under the assumptions (4.1) - (4.3) for the general case when primaries are of unequal size and neither $\rho = \sigma_\alpha^2/\sigma_\varepsilon^2$ nor $\sigma_\alpha^2$ and $\sigma_\varepsilon^2$ are known. The notation used here is consistent with that of Section 3. In addition, $\underline{X}$ is a $(M_o \times 1)$ vector of the $x_{ij}$'s in the finite population, and $\underline{X}_s$ is a $(m_o \times 1)$ vector of the $x_{ij}$'s in the sample.

## 4.2 Estimation and Variance Formulas for the Finite Population Mean

The finite population can be expressed as

$$y_{ij} = a_i + bx_{ij} + e_{ij} \ , \qquad i = 1, \ldots, N \ , \quad j = 1, \ldots, M_i, \tag{4.4}$$

where

$$b = \frac{\sum_{i}^{N}\sum_{j}^{M_i}(x_{ij}-\bar{x}_i)y_{ij}}{\sum_{i}^{N}\sum_{j}^{M_i}(x_{ij}-\bar{x}_i)^2} \tag{4.5}$$

and

$$a_i = \frac{\sum_{j}^{M_i} y_{ij}}{M_i} - b\frac{\sum_{j}^{M_i} x_{ij}}{M_i}, \qquad i = 1, \ldots, N , \tag{4.6}$$

are the least square coefficients for the regression of $y_{ij}$ on $x_{ij}$.

The mean of the finite population can be represented as

$$\bar{Y} = \frac{\sum_{i}^{N} M_i\bar{Y}_i}{M_o}$$

$$= \frac{\sum_{i}^{N} M_i(a_i+b\bar{X}_i)}{M_o} . \tag{4.7}$$

An estimator of $\bar{Y}$ can therefore be constructed by estimating b and $a_i$ from the sample. Performing the regression of $y_{ij}$ on $x_{ij}$ for the sample yields

$$\hat{b} = \frac{\sum_{i}^{n}\sum_{j}^{m_i}(x_{ij}-\bar{x}_i)y_{ij}}{\sum_{i}^{n}\sum_{j}^{m_i}(x_{ij}-\bar{x}_i)^2} \tag{4.8}$$

and

$$\hat{a}_i = \bar{y}_i - \hat{b}\bar{x}_i , \qquad i = 1, \ldots, n , \tag{4.9}$$

where

$$\bar{y}_i = \frac{\sum_j^{m_i} y_{ij}}{m_i} \tag{4.10}$$

and

$$\bar{x}_i = \frac{\sum_j^{m_i} x_{ij}}{m_i} . \tag{4.11}$$

As an estimator of $a_i$ for the primaries not selected in the sample, the knowledge that the conditional expectation of $a_i$ given any $\underline{X}$ is $\mu$, suggests using an unbiased estimator of $\mu$. Following Section 3.2, let the estimator for $a_i$ when the $i^{th}$ primary is not selected in the sample be

$$\hat{a}_i = \bar{y}_a - \hat{b}\bar{x}_a \tag{4.12}$$

where

$$\bar{y}_a = \sum_i^n k_i m_i \bar{y}_i , \tag{4.13}$$

$$\bar{x}_a = \sum_i^n k_i m_i \bar{x}_i , \tag{4.14}$$

and

$$k_i = \frac{m_o - m_i}{m_o^2 - \sum_i^n m_i^2} . \tag{4.15}$$

Substituting $\hat{b}$ and $\hat{a}_i$ from (4.8), (4.9), and (4.12) for b and $a_i$ in (4.7) gives

$$\hat{\bar{Y}}_R = \pi\bar{y}_r + (1 - \pi)\bar{y}_a + \hat{b}[\bar{X} - (\pi\bar{x}_r + (1 - \pi)\bar{x}_a)] \qquad (4.16)$$

where

$$\pi = \frac{\sum_i^n M_i}{M_o} , \qquad (4.17)$$

$$\bar{y}_r = \sum_i^n M_i\bar{y}_i / \sum_i^n M_i , \qquad (4.18)$$

$$\bar{x}_r = \sum_i^n M_i\bar{x}_i / \sum_i^n M_i , \qquad (4.19)$$

and

$$\bar{X} = \frac{\sum_i^N \sum_j^{M_i} x_{ij}}{M_o} . \qquad (4.20)$$

Obviously, to use this estimator, $\bar{X}$ must be known. It is of interest to note that when $n = N$, $\pi = 1$, and $\hat{\bar{Y}}_R$ simplifies to the classical combined regression estimator for stratified sampling, namely,

$$\hat{\bar{Y}}_R = \bar{y}_r + \hat{b}(\bar{X} - \bar{x}_r) . \qquad (4.21)$$

The difference $\bar{Y} - \hat{\bar{Y}}_R$ can be expressed as

$$\bar{Y} - \hat{\bar{Y}}_R = \frac{1}{M_o}\left\{\sum_i^{N-n} Y_i + \sum_i^n\left[Y_i - y_i\left(\frac{M_i}{m_i} + (1 - \pi)M_o k_i\right)\right]\right\}$$

$$- \hat{b}[\bar{X} - (\pi\bar{x}_r + (1 - \pi)\bar{x}_a)] \qquad (4.22)$$

where

$$Y_i = \sum_{j}^{M_i} y_{ij} \qquad (4.23)$$

and

$$y_i = \sum_{j}^{m_i} y_{ij} \; . \qquad (4.24)$$

Noting that

$$E(Y_i | \underline{X}, \underline{X}_s) = M_i \mu + \beta \sum_{j}^{M_i} x_{ij} \; , \qquad (4.25)$$

$$E(y_i | \underline{X}, \underline{X}_s) = m_i \mu + \beta \sum_{j}^{m_i} x_{ij} \; , \qquad (4.26)$$

and

$$E(\hat{b} | \underline{X}, \underline{X}_s) = \beta \; , \qquad (4.27)$$

it follows that

$$E(\bar{Y} - \hat{\bar{Y}}_R | \underline{X}, \underline{X}_s) = 0 \; . \qquad (4.28)$$

Also,

$$V(Y_i | \underline{X}, \underline{X}_s) = M_i^2 \sigma_\alpha^2 + M_i \sigma_\varepsilon^2 \; , \qquad (4.29)$$

$$V(y_i | \underline{X}, \underline{X}_s) = m_i^2 \sigma_\alpha^2 + m_i \sigma_\varepsilon^2 \; , \qquad (4.30)$$

$$\mathrm{Cov}(Y_i , \; y_i | \underline{X}, \underline{X}_s) = m_i M_i \sigma_\alpha^2 + m_i \sigma_\varepsilon^2 \; , \qquad (4.31)$$

$$\mathrm{Cov}(Y_i , \; \hat{b} | \underline{X}, \underline{X}_s) = 0 \; , \qquad (4.32)$$

$$\mathrm{Cov}(y_i , \; \hat{b} | \underline{X}, \underline{X}_s) = 0 \; , \qquad (4.33)$$

and

$$V(\hat{b}\,|\,\underline{X},\ \underline{X}_s) = \frac{\sigma_\varepsilon^2}{s_x^2} \tag{4.34}$$

where

$$s_x^2 = \sum_i^n \sum_j^{m_i} (x_{ij} - \bar{x}_i)^2 \ . \tag{4.35}$$

It follows that

$$\begin{aligned} V(\bar{Y} - \hat{\bar{Y}}_R\,|\,\underline{X},\ \underline{X}_s) &= \frac{\sigma_\alpha^2}{M_o^2}\left\{\sum_i^{N-n} M_i^2 + M_o^2(1 - \pi)^2 \sum_i^n k_i^2 m_i^2\right\} \\ &+ \frac{\sigma_\varepsilon^2}{M_o^2}\Biggl\{\sum_i^n \frac{M_i^2}{m_i} + M_o(1 - 2\pi + (1 - \pi)[M_o(1 - \pi)\sum_i^n k_i^2 m_i \\ &+ 2\sum_i^n k_i(M_i - m_i)])\Biggr\} \\ &+ \sigma_\varepsilon^2\left\{\frac{[\bar{X} - (\pi\bar{x}_r + (1-\pi)\bar{x}_a)]^2}{s_x^2}\right\} \ . \end{aligned} \tag{4.36}$$

If all $m_i = m$, this reduces to

$$V(\bar{Y} - \hat{\bar{Y}}_R | \underline{X}, \underline{X}_s) = \frac{\sigma_\alpha^2}{n} \left\{ \frac{n \sum_i^{N-n} M_i^2}{M_o^2} + (1 - \pi)^2 \right\}$$

$$+ \frac{\sigma_\varepsilon^2}{mn} \left\{ \frac{n}{M_o} \left( \frac{\sum_i^n M_i^2}{M_o} - m \right) + (1 - \pi^2) \right\}$$

$$+ \sigma_\varepsilon^2 \frac{[\bar{X} - (\pi \bar{x}_r + (1-\pi) \bar{x}_a)]^2}{s_x^2} \quad . \qquad (4.37)$$

If in addition, all $M_i = M$,

$$V(\bar{Y} - \hat{\bar{Y}}_R | \underline{X}, \underline{X}_s) = \frac{\sigma_\alpha^2}{n} (1 - \frac{n}{N}) + \frac{\sigma_\varepsilon^2}{mn} (1 - \frac{mn}{MN})$$

$$+ \sigma_\varepsilon^2 \frac{\left( \bar{X} - \frac{\sum\sum_{ij}^{nm} x_{ij}}{mn} \right)^2}{s_x^2} \quad . \qquad (4.38)$$

## 4.3 Confidence Intervals on $\bar{Y}$

In constructing a confidence interval on $\bar{Y}$, it is assumed that all $m_i = m$. Through the use of Theorem 2.1 of Section 2.2,

$$g = g_1 s_b^2 + g_2 s_w^2 \qquad (4.39)$$

where

$$g_1 = \frac{1}{m}\left(\frac{\sum_{i}^{N-n} M_i^2}{M_o^2} + \frac{(1-\pi)^2}{n}\right) \qquad (4.40)$$

and

$$g_2 = \left(\frac{\sum_{i}^{n} M_i^2 - \sum_{i}^{N-n} M_i^2}{M_o^2 m}\right) - \frac{1}{M_o} + \frac{2\pi(1-\pi)}{mn}$$

$$+ \frac{[\bar{X} - (\pi\bar{x}_r + (1-\pi)\bar{x}_a)]^2}{s_x^2} , \qquad (4.41)$$

is an unbiased estimator for $V(\bar{Y} - \hat{\hat{Y}}_R | \underline{X}, \underline{X}_s)$. In addition, the conditional distribution of $(\bar{Y} - \hat{\hat{Y}}_R)/\sqrt{g}$ given $\underline{X}$ and $\underline{X}_s$ is an approximate t-distribution with n' degrees of freedom where

$$n' = \frac{(g_1\sigma_1^2 + g_2\sigma_2^2)^2}{g_1^2\sigma_1^4/(n-1) + g_2^2\sigma_2^4/(n)(m-1)} , \qquad (4.42)$$

$$\sigma_1^2 = \sigma_\varepsilon^2 + m\sigma_\alpha^2 , \qquad (4.43)$$

and

$$\sigma_2^2 = \sigma_\varepsilon^2 . \qquad (4.44)$$

Hence, only $\bar{X}$ and the observed $x_{ij}$'s in the sample are needed to completely specify the t-distribution, and an approximate $100(1-\alpha)\%$ confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}}_R \pm t_{\alpha/2;n'}\sqrt{g}\ ] . \qquad (4.45)$$

Note that the confidence interval (4.45) depends on $\bar{X}$ and $X_s$ since both g and n' depend on $\bar{X}$ and $X_s$. However, since the confidence level, 100(1-α)%, does not depend on either $\bar{X}$ or $X_s$, the above confidence interval procedure produces intervals containing $\bar{Y}$ 100(1-α)% of the time.

An exact confidence interval for $\bar{Y}$ cannot be constructed using the technique of Section 2.2. If $u_i$ were defined to be

$$u_i = c_1\bar{y}_i + c_2d_i \qquad (4.46)$$

where

$$d_i = \sum_j^m \ell_{ij}y_{ij} \qquad (4.47)$$

and

$$\sum_j^m \ell_{ij} = 0 , \qquad (4.48)$$

then

$$E(u_i|\underline{X}, \underline{X}_s) = c_1\mu + \beta(c_1\bar{x}_i + c_2\sum_j^m \ell_{ij}x_{ij}) \qquad (4.49)$$

and the $u_i$'s would not have the same mean. Therefore $\sum_i^n(u_i - \bar{u})^2$ given $\underline{X}$ and $\underline{X}_s$ would not be a central chi-square, and hence an exact t-distribution would not be available.

### 4.4 Two-Stage Sampling When Elements are Related to Primary Size

An important special case of the model in (4.1) is when $x_{ij} = M_i$, i.e., an element in the population is related to the size of its primary. A similar model has been considered by Cochran [1963] to study the behavior of estimators for single-stage cluster sampling designs. The super-population is now defined as

$$y_{ij} = \mu + \alpha_i + \beta M_i + \varepsilon_{ij} \tag{4.50}$$

where

$$\alpha_i \sim N(0, \sigma_\alpha^2) \ , \tag{4.51}$$

$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2) \ , \tag{4.52}$$

$M_i$, $\alpha_i$, and $\varepsilon_{ij}$ are all independent, and $\mu$ and $\beta$ are constants. The finite population can be expressed as

$$y_{ij} = a + bM_i + e_{ij} \tag{4.53}$$

where

$$a = \bar{Y} - b \frac{\sum_i^N M_i^2}{M_o} \tag{4.54}$$

and

$$b = \frac{\sum_i^N \sum_j^{M_i} M_i y_{ij} - \bar{Y} \sum_i^N M_i^2}{\sum_i^N M_i^3 - \frac{\left(\sum_i^N M_i^2\right)^2}{M_o}} \ . \tag{4.55}$$

The finite population mean is now

$$\bar{Y} = a + b\,\frac{\sum_{i}^{N} M_i^2}{M_o} \,. \tag{4.56}$$

A logical estimator for $\bar{Y}$ is

$$\hat{\bar{Y}}_M = \hat{a} + \hat{b}\,\frac{\sum_{i}^{N} M_i^2}{M_o} \tag{4.57}$$

where

$$\hat{a} = \bar{y} - \hat{b}\,\frac{\sum_{i}^{n} M_i m_i}{m_o} \tag{4.58}$$

and

$$\hat{b} = \frac{\sum_{i}^{n} M_i m_i \bar{y}_i - \bar{y}(\sum_{i}^{n} M_i m_i)}{\sum_{i}^{n} M_i^2 m_i - (\sum_{i}^{n} M_i m_i)^2/m_o} \tag{4.59}$$

are the least square coefficients for the sample regression of $y_{ij}$ on $M_i$. Simplifying (4.57) gives

$$\hat{\bar{Y}}_M = \bar{y} + \hat{b}\left(\frac{\sum_{i}^{N} M_i^2}{M_o} - \frac{\sum_{i}^{n} M_i m_i}{m_o}\right) . \tag{4.60}$$

Noting that

$$E(\bar{y}|M_i) = \mu + \beta\,\frac{\sum_{i}^{n} M_i m_i}{m_o} \,, \tag{4.61}$$

$$E(\hat{b}|M_i) = \beta \,, \tag{4.62}$$

and

$$E(\bar{Y}|M_i) = \mu + \beta\,\frac{\sum_{i}^{N} M_i^2}{M_o} \,, \tag{4.63}$$

it follows that

$$E(\bar{Y} - \hat{\bar{Y}}_M | M_i) = 0 \ . \tag{4.64}$$

In addition,

$$V(\bar{Y} - \hat{\bar{Y}}_M | M_i) = \sigma_\alpha^2 \left\{ \frac{\sum_i^{N-n} M_i^2}{M_o^2} + \sum_i^n \left( \frac{M_i}{M_o} - \frac{m_i(1-C)}{m_o} - \frac{M_i m_i C}{\sum_i^n M_i m_i} \right)^2 \right\}$$

$$+ \sigma_\varepsilon^2 \left\{ 1 - \frac{m_o}{M_o} + \sum_i^n m_i \left( \frac{1}{M_o} - \frac{(1-C)}{m_o} - \frac{M_i C}{\sum_i^n M_i m_i} \right)^2 \right\} \tag{4.65}$$

where

$$C = \frac{\left(\sum_i^n M_i m_i\right) \left( \frac{\sum_i^N M_i^2}{M_o} - \frac{\sum_i^n M_i m_i}{m_o} \right)}{\sum_i^n M_i^2 m_i - \frac{\left(\sum_i^n M_i m_i\right)^2}{m_o}} \ . \tag{4.66}$$

If all $m_i = m$, an approximate $100(1-\alpha)\%$ confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}}_M \pm t_{\alpha/2;n'} \sqrt{g} \ ] \tag{4.67}$$

where

$$g = g_1 s_b^2 + g_2 s_w^2 , \tag{4.68}$$

$$g_1 = \frac{1}{m} \left\{ \frac{\sum_i^{N-n} M_i^2}{M_o^2} + \frac{(1-\pi)^2}{n} + \frac{A^2}{B} \right\} , \tag{4.69}$$

$$g_2 = \frac{1}{mM_o} \left[ \frac{\sum_i^{n} M_i^2}{M_o} - m + 2A \right] + \frac{(1-\pi^2)}{mn} + \frac{A^2}{mB} - g_1 , \tag{4.70}$$

$$A = \frac{\sum_i^{N-n} M_i^2}{M_o} - \frac{(1-\pi)(\pi)M_o}{n} , \tag{4.71}$$

$$B = \sum_i^{n} M_i^2 - \frac{\pi^2 M_o^2}{n} , \tag{4.72}$$

and n' is defined in (4.42).

## 4.5 Robustness of $\hat{\bar{Y}}_R$ to Model Breakdown

The robustness of $\hat{\bar{Y}}_R$ to model breakdown is now discussed. Notice that $\hat{\bar{Y}}_R$ can be written as

$$\hat{\bar{Y}}_U + \hat{b}[\bar{X} - (\pi\bar{x}_r + (1 - \pi)\bar{x}_a)] \tag{4.73}$$

where $\hat{\bar{Y}}_U$ is defined in (3.26). As was shown in Section 3.4, for a given finite population $\hat{\bar{Y}}_U$ is a biased estimator of $\bar{Y}$. It therefore follows that $\hat{\bar{Y}}_R$ is also biased for $\bar{Y}$ given a fixed finite population.

Furthermore, in the special case of stratified sampling where $\hat{\bar{Y}}_R$ is the classical combined regression estimator, $\hat{\bar{Y}}_R$ remains biased even though $\hat{\bar{Y}}_U$ is now an unbiased estimator of $\bar{Y}$ (see, e.g., Cochran [1963, p. 202]).

# 5. EXTENSIONS AND CONCLUSIONS

## 5.1 Sampling with More Than Two Stages

### 5.1.1 A p-stage sampling design

The methodology developed in this report can be extended in a straightforward manner to higher levels than the two-stage designs considered previously. For a p-stage design, the linear model describing the super-population is assumed to be the random p-fold nested classification model. It is further assumed that all random factors are normally distributed with mean zero and constant variance, and that all factors are independent.

If all sampling units are of equal size, the least squares technique used in Section 2.1 can be used to construct estimators for the finite population parameters. If sampling units are of unequal size, as in Section 3.2, estimators can be constructed by using the knowledge that non-sampled units are from a distribution defined by the super-population model.

If equal samples per sampling unit are selected, an approximate confidence interval on $\bar{Y}$ can be constructed by using the following result from variance components analysis (see, e.g., Graybill [1961, p. 347]).

Theorem 5.1: Let the random p-fold nested classification model hold for the super-population and let all random variables in the model be independent and normally distributed. Let the mean square for the $i^{th}$ factor in the model be denoted by $s_i^2$. Let $s_i^2$ have $n_i$ degrees of freedom and let $E(s_i^2) = \sigma_i^2$. Then $v_i = n_i s_i^2/\sigma_i^2 \sim \chi^2_{(n_i)}$ and $v_1, \ldots, v_p$ are mutually independent.

Using Theorem 2.1 of Section 2.2, it follows that $g = \sum_i^p g_i s_i^2$ is an unbiased estimator of $\gamma = \sum_i^p g_i \sigma_i^2$, and

$$\frac{n'g}{\gamma} \overset{\cdot}{\sim} \chi^2_{(n')} \tag{5.1}$$

where

$$n' = \frac{\left(\sum_i^p g_i \sigma_i^2\right)^2}{\sum_i^p (g_i^2 \sigma_i^4 / n_i)} \; . \tag{5.2}$$

By selecting the $g_i$ in such a manner that $\gamma = V(\bar{Y} - \hat{\bar{Y}})$, an approximate $100(1-\alpha)\%$ confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}} \pm t_{\alpha/2;n'} \sqrt{g}\ ] \; . \tag{5.3}$$

An exact confidence interval on $\bar{Y}$ may also be constructed by considering appropriate linear contrasts of the observations. An example for the three-stage design is given in the following section.

### 5.1.2 Three-stage sampling with equal sizes and samples

For the three-stage sampling design let

N = number of primaries,

M = number of secondaries/primary,

L = number of tertiaries/secondary,

n = number of sampled primaries,

m = number of sampled secondaries/sampled primary, and

$\ell$ = number of sampled tertiaries/sampled secondary.

The three-fold nested classification model describing the superpopulation is

$$y_{ijk} = \mu + \eta_{ijk} \tag{5.4}$$

where

$$E(\eta_{ijk}) = 0 \tag{5.5}$$

and

$$\begin{aligned} E(\eta_{ijk} \; \eta_{i'j'k'}) &= \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2 , & & i=i', \ j=j', \ k=k', \\ &= \sigma_\alpha^2 + \sigma_\beta^2 , & & i=i', \ j=j', \ k\neq k', \\ &= \sigma_\alpha^2 , & & i=i', \ j\neq j', \\ &= 0 , & & i\neq i' . \end{aligned} \tag{5.6}$$

Using the methodology of Section 2.1,

$$\hat{\hat{Y}} = \frac{\overset{nm\ell}{\underset{ijk}{\Sigma\Sigma\Sigma}} y_{ijk}}{\ell mn} \tag{5.7}$$

and

$$V(\bar{Y} - \hat{\hat{Y}}) = \frac{\sigma_\alpha^2}{n}(1 - \frac{n}{N}) + \frac{\sigma_\beta^2}{mn}(1 - \frac{mn}{MN}) + \frac{\sigma_\varepsilon^2}{\ell mn}(1 - \frac{\ell mn}{LMN}) \quad . \tag{5.8}$$

Alternatively, $V(\bar{Y} - \hat{\hat{Y}})$ can be expressed as

$$V(\bar{Y} - \hat{\bar{Y}}) = g_1\sigma_1^2 + g_2\sigma_2^2 + g_3\sigma_3^2 \tag{5.9}$$

where

$$g_1 = (1 - \frac{n}{N}) \frac{1}{\ell mn} , \tag{5.10}$$

$$g_2 = (1 - \frac{m}{M}) \frac{1}{\ell mN} , \tag{5.11}$$

$$g_3 = (1 - \frac{\ell}{L}) \frac{1}{\ell MN} , \tag{5.12}$$

$$\sigma_1^2 = \ell m\sigma_\alpha^2 + \ell\sigma_\beta^2 + \sigma_\varepsilon^2 , \tag{5.13}$$

$$\sigma_2^2 = \ell\sigma_\beta^2 + \sigma_\varepsilon^2 , \tag{5.14}$$

and

$$\sigma_3^2 = \sigma_\varepsilon^2 . \tag{5.15}$$

Using Theorem 5.1, an unbiased estimator for $V(\bar{Y} - \hat{\bar{Y}})$ is

$$g = g_1s_1^2 + g_2s_2^2 + g_3s_3^2 \tag{5.16}$$

where

$$s_1^2 = \frac{\sum\limits_{ijk}^{nm\ell}\sum\sum(\bar{y}_i - \bar{y})^2}{n-1} , \tag{5.17}$$

$$s_2^2 = \frac{\sum\limits_{ijk}^{nm\ell}\sum\sum(\bar{y}_{ij} - \bar{y}_i)^2}{n(m-1)} , \tag{5.18}$$

$$s_3^2 = \frac{\sum\limits_{ijk}^{nm\ell}\sum\sum(y_{ijk} - \bar{y}_{ij})^2}{nm(\ell-1)} , \tag{5.19}$$

$$\bar{y}_i = \frac{\sum_{jk}^{m\ell}\sum y_{ijk}}{\ell m} , \qquad (5.20)$$

$$\bar{y}_{ij} = \frac{\sum_{k}^{\ell} y_{ijk}}{\ell} , \qquad (5.21)$$

and

$$\bar{y} = \frac{\sum_{ijk}^{nm\ell}\sum\sum y_{ijk}}{\ell mn} . \qquad (5.22)$$

An approximate confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}} \pm t_{\alpha/2;n'}\sqrt{g}\ ] \qquad (5.23)$$

where

$$n' = \frac{(g_1^2\sigma_1^2+g_2^2\sigma_2^2+g_2^2\sigma_3^2)^2}{g_1^2\sigma_1^4/(n-1)+g_2^2\sigma_2^4/(n(m-1))+g_3^2\sigma_3^4/(nm(\ell-1))} . \qquad (5.24)$$

An exact confidence interval can be constructed by considering the random variable

$$u_{ij} = c_1\bar{y}_i + c_2 d_i + c_3 f_{ij} \qquad (5.25)$$

where

$$d_i = \sum_{j}^{m} r_{ij}\bar{y}_{ij} , \qquad (5.26)$$

$$\sum_{j}^{m} r_{ij} = 0 , \qquad (5.27)$$

$$f_{ij} = \sum_{k}^{\ell} h_{ijk} y_{ijk} , \qquad (5.28)$$

$$\sum_{k}^{\ell} h_{ijk} = 0 \ , \tag{5.29}$$

$$c_4 = \sum_{j}^{m} r_{ij}^2 \ , \tag{5.30}$$

$$c_5 = \sum_{k}^{\ell} h_{ijk}^2 \ , \tag{5.31}$$

and the $r_{ij}$'s, $h_{ijk}$'s, $c_1$, $c_2$, $c_3$, $c_4$, and $c_5$ are constants. Assuming model (5.4) - (5.6), it follows that for all i and j

$$\bar{y}_i \sim N(\mu,\ \sigma_\alpha^2 + \sigma_\beta^2/m + \sigma_\varepsilon^2/\ell m) \ , \tag{5.32}$$

$$d_i \sim N(0,\ c_4(\sigma_\beta^2 + \sigma_\varepsilon^2/\ell) \ , \tag{5.33}$$

$$f_{ij} \sim N(0,\ c_5\sigma_\varepsilon^2) \ , \tag{5.34}$$

$$\mathrm{Cov}(\bar{y}_i,\ d_i) = 0 \ , \tag{5.35}$$

$$\mathrm{Cov}(\bar{y}_i,\ f_{ij}) = 0 \ , \tag{5.36}$$

and

$$\mathrm{Cov}(d_i,\ f_{ij}) = 0 \ . \tag{5.37}$$

Equating $V(u_{ij})$ to $V(\bar{Y} - \hat{\bar{Y}})$ and solving for $c_1^2$, $c_2^2c_4$, and $c_3^2c_5$ yields

$$c_1^2 = \frac{1}{n}(1 - \frac{n}{N}) \ , \tag{5.38}$$

$$c_2^2c_4 = \frac{1}{Nm}(1 - \frac{m}{M}) \ , \tag{5.39}$$

and

$$c_3^2c_5 = \frac{1}{NM\ell}(1 - \frac{\ell}{L}) \ . \tag{5.40}$$

Using the results of Section 2.2, and noting that

$$\frac{\sum_{i}^{n}\sum_{j}^{m}(u_{ij}-\bar{u}_i)^2}{V(u_{ij})} \sim \chi^2_{n(m-1)} \text{ ,} \tag{5.41}$$

an exact 100(1-α)% confidence interval on $\bar{Y}$ is

$$[\hat{\bar{Y}} \pm t_{\alpha/2;n(m-1)}\sqrt{g_e}] \tag{5.42}$$

where

$$g_e = \frac{\sum_{i}^{n}\sum_{j}^{m}(u_{ij}-\bar{u}_i)^2}{n(m-1)} \tag{5.43}$$

and

$$u_{ij} = \sqrt{\frac{1}{n}(1-\frac{n}{N})}\ \bar{y}_i + \sqrt{\frac{1}{Nm}(1-\frac{m}{M})}\ \frac{\sum_{j}^{m} r_{ij}\bar{y}_{ij}}{\sqrt{\sum_{j}^{m} r_{ij}^2}}$$

$$+ \sqrt{\frac{1}{NM\ell}(1-\frac{\ell}{L})}\ \frac{\sum_{k}^{\ell} h_{ijk}y_{ijk}}{\sqrt{\sum_{k}^{\ell} h_{ijk}^2}} \text{ .} \tag{5.44}$$

## 5.2 Concluding Remarks and Future Research

In many practical sampling situations the rationale for the super-population model is more realistic than that of classical frequentist theory. If the population of interest changes over time, repeated samples from this population are in fact repeated samples from a super-population and not a fixed finite population. Trueblood and Cyert [1957] offer an example of a two-stage sample design used to confirm accounts receivable in a department store. Over a period of time, the account ledgers change, and the notion of a super-population generating a fixed set of ledgers at a given point in time seems to be an improvement over the classical assumptions. The results developed in this report are appropriate when sampling from such a population.

Possibilities for future research in the super-population theory of survey sampling include

(i) the study of ratio estimators in both single and multi-stage designs,

(ii) an Empirical Bayes approach for estimation of the finite population parameters,

(iii) the further study of estimation of finite population parameters in surveys with more than two stages and sampling units of unequal size, and

(iv) an extension of the results of Section 4 to the multivariate case where $x_{ij}$ is a vector rather than a single variable.

Royall and Herson [1973a] have considered a special case of a ratio estimator in single stage designs where the super-population model is a polynomial regression and have found it to be the BLUE of the finite population total. Further research into the behavior of ratio estimators in multi-stage surveys may produce similar results.

Since super-population theory implies the parameters of the finite population are truly random variables, it might be worthwhile to consider an Empirical Bayes approach to the problem of estimating the finite population parameters. An Empirical Bayes approach would utilize information from previous surveys of the population to estimate the current finite population parameters.

In extending the methodology developed in this report to more than two stages when sampling units are of unequal size, various estimators for non-sampled elements are plausible and these alternatives should perhaps be studied.

An appropriate super-population model for extending the results of Section 4 to the multivariate case is

$$y_{ij} = \mu + \alpha_i + \underline{x}_{ij}^T \underline{\beta} + \varepsilon_{ij} \tag{5.45}$$

where $\underline{x}_{ij}$ is a vector of variables and $\underline{\beta}$ is a vector of constants. Using matrix notation and the arguments of Section 4, estimators for the multivariate case could be developed.

In conclusion, a super-population theory to survey sampling is quite realistic and brings survey sampling more into the mainstream of statistical inference than does the classical theory depicted as Case 1 in Table 1 (p. 7). Future research will hopefully lead to a more comprehensive theory which will serve as the framework for all sample surveys.

# REFERENCES

Barnard, G. A. [1969]. Summary Remarks. In Johnson, N. L., and Smith, H., New Developments in Survey Sampling, Wiley, New York.

Cochran, W. G. [1939]. The Use of Analysis of Variance in Enumeration by Sampling. J. Amer. Statist. Assn. 34, 492-510.

Cochran, W. G. [1946]. Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations. Ann. Math. Statist. 17, 164-177.

Cochran, W. G. [1963]. Sampling Techniques, Wiley, New York.

Deming, W. E. [1950]. Some Theory of Sampling, Wiley, New York.

Ericson, W. A. [1969a]. Subjective Bayesian Models in Sampling Finite Populations I. J. Royal Statist. Soc. B 31, 195-234.

Ericson, W. A. [1969b]. Subjective Bayesian Models in Sampling Finite Populations: Stratification. In Johnson, N. L., and Smith, H., New Developments in Survey Sampling, Wiley, New York.

Fuller, W. A. [1973]. Regression Analysis for Sample Surveys, paper presented at the Vienna meeting of the International Institute of Survey Statisticians.

Gnedenko, B. V. [1963]. The Theory of Probability, Chelsea, New York.

Godambe, V. P. [1955]. A Unified Theory of Sampling from Finite Populations. J. Royal Statist. Soc. B 17, 269-278.

Godambe, V. P. [1966]. A New Approach to Sampling from Finite Populations-II. J. Royal Statist. Soc. B 28, 310-328.

Godambe, V. P. [1969]. Some Aspects of the Theoretical Developments in Survey Sampling. In Johnson, N. L., and Smith, J., New Developments in Survey Sampling, Wiley, New York.

Godambe, V. P. [1970]. Foundations of Survey Sampling. The Amer. Statist. 24, 33-38.

Godambe, V. P. and Sprott, D. A. [1971]. Foundations of Statistical Inference, Holt, Rinehart, and Winston, Toronto.

Graybill, F. A. [1961]. An Introduction to Linear Statistical Models, Volume I, Mc-Graw Hill, New York.

Hartley, H. O., and Rao, J. N. K. [1968]. A New Estimation Theory for Sample Surveys. Biometrika 55, 547-557.

Hartley, H. O., and Rao, J. N. K. [1969]. A New Estimation Theory for Sample Surveys, II. In Johnson, N. L., and Smith, H., *New Developments in Survey Sampling*, Wiley, New York.

Hartley, H. O., and Sielken, R. L. [1975]. A "Super-Population Viewpoint" for Finite Population Sampling. *Biometrics 31*, 411-422.

Harville, D. A. [1969]. Expression of Variance Component Estimators as Linear Combinations of Independent Non-Central Chi-Square Variates. *Ann. Math. Statist. 40*, 2189-2194.

Henderson, C. R. [1953]. Estimation of Variance and Covariance Components. *Biometrics 9*, 226-252.

Hogg, R. V., and Craig, A. T. [1970]. *Introduction to Mathematical Statistics*, Macmillan Company, New York.

Johnson, N. L., and Smith, H. [1969]. *New Developments in Survey Sampling*, Wiley, New York.

Kish, L. [1952]. A Two-Stage Sample of a City. *Amer. Sociological Review 17*, 761-769.

Kish, L. [1965]. *Survey Sampling*, Wiley, New York.

Koch, G. G. [1967]. A Procedure to Estimate the Population Mean in Random Effects Models. *Technometrics 9*, 577-585.

Konijn, H. S. [1962]. Regression Analysis in Sample Surveys. *J. Amer. Statist. Assn. 57*, 590-606.

Mendenhall, W., Ott, L., and Scheaffer, R. L. [1971]. *Elementary Survey Sampling*, Wadsworth, Belmont, Calif.

Raj, D. [1958]. On the Relative Accuracy of Some Sampling Techniques. *J. Amer. Statist. Assn. 53*, 98-101.

Raj, D. [1968]. *Sampling Theory*, McGraw-Hill, New York.

Rao, J. N. K. [1971]. *Some Thoughts on the Foundations of Survey Sampling*, Tech. Report No. 10, The University of Manitoba.

Rao, J. N. K. [1973]. *On the Foundations of Survey Sampling*, paper presented at International Symposium on Statistical Design and Linear Models, Colorado State University, March 1973.

Royall, R. M. [1968]. An Old Approach to Finite Population Sampling Theory. *J. Amer. Statist. Assn. 63*, 1269-1279.

Royall, R. M. [1970]. On Finite Population Sampling Theory Under Certain Linear Regression Models. *Biometrika 57*, 377-387.

Royall, R. M., and Herson, J. [1973a]. Robust Estimation in Finite Populations I. *J. Amer. Statist. Assn. 68*, 880-889.

Royall, R. M., and Herson, J. [1973b]. Robust Estimation in Finite Populations II: Stratification on a Size Variable. *J. Amer. Statist. Assn. 68*, 890-893.

Scott, Alastair, and Smith, T. M. F. [1969]. Estimation in Multi-Stage Surveys. *J. Amer. Statist. Assn. 64*, 830-840.

Searle, S. R. [1971a]. *Linear Models*, Wiley, New York.

Searle, S. R. [1971b]. Topics in Variance Components Estimation. *Biometrics 27*, 1-76.

Sedransk, J. [1965]. Analytical Surveys with Cluster Sampling. *J. Royal Statist. Soc. B 27*, 264-278.

Sukhatme, P. V. [1947]. The Problem of Plot Size in Large-Scale Yield Surveys. *J. Amer. Statist. Assn. 42*, 297-310.

Trueblood, R. M., and Cyert, R. M. [1957]. *Sampling Techniques in Accounting*, Prentice-Hall, Englewood Cliffs, N. J.

## APPENDIX A: A NUMERICAL EXAMPLE FOR TWO-STAGE SAMPLING WITH EQUAL SIZES AND SAMPLES

The following example is taken from Mendenhall et al. [1971, p. 186] with the modification that all primaries are of equal size.

A forester wants to estimate the total number of trees in a certain county which are infected with a particular disease. A two-stage sampling design is used with the county being divided into ten primary units which are further subdivided into fifteen secondaries of approximately equal size. A sample of four primaries and six secondaries per sampled primary is taken. The survey results are given in Table 3.

TABLE 3

Results of Forester's Survey

| Area | Number of Infected Trees per Plot ($y_{ij}$) |
|---|---|
| 1 | 15, 14, 21, 13, 9, 10 |
| 2 | 4, 6, 10, 9, 8, 5 |
| 3 | 10, 11, 14, 10, 9, 15 |
| 4 | 8, 3, 4, 1, 2, 5 |

For this problem N = 10, M = 15, n = 4, and m = 6. From (2.14)

$$\hat{\bar{Y}}_E = \frac{216}{24} = 9 \, .$$

Also, (2.28), (2.29), and (2.35) give

$$s_b^2 = 117.444 ,$$

$$s_w^2 = \quad 8.983 ,$$

and

$$g = 3.026 .$$

The degrees of freedom given in (2.37) associated with the approximate confidence interval are

$$n' = 3.186 .$$

Since n' is not an integer, it is rounded down to 3, and a conservative 95% confidence interval on $\bar{Y}$ from (2.43) is

$$(3.468,\ 14.532) .$$

To compute the exact confidence interval on $\bar{Y}$ the $\ell_{ij}$ are selected to be

$$\begin{aligned} \ell_{ij} &= -1 , \qquad j = 1, 2, 3 , \\ &= +1 , \qquad j = 4, 5, 6 , \end{aligned}$$

for all sampled primaries. Then from (2.46)

$$d_1 = -18 ,$$

$$d_2 = \ +2 ,$$

$$d_3 = \ -1 ,$$

and

$$d_4 = \ -7 .$$

Also, from (2.55), (2.56), and (2.59),

$$c_1^2 = .15 ,$$

$$c_2^2 c_3 = .01 ,$$

and

$$g_e = 2.493 .$$

The exact 95% confidence interval on $\bar{Y}$ from (2.64) is

$$(3.979, 14.02) .$$

# VITA

Richard K. Burdick was born July 12, 1950, in Sterling, Colorado. His parents, Mr. and Mrs. Keith Burdick, live at 2124 Rainbow, Laramie, Wyoming. Rick grew up in Laramie, and graduated from Laramie Senior High School in May, 1968. After attending Colorado School of Mines for two years, he transferred to the University of Wyoming where he received the B.S. degree in statistics in May, 1972. He has studied since September, 1972, at Texas A&M University where he received the Master of Statistics degree in December, 1973, and anticipates receiving the Doctor of Philosophy degree in statistics in August, 1976.

The typist for this dissertation was Karon Freyer.